Hybrid-time encoded speech.

Singh, Amarjit

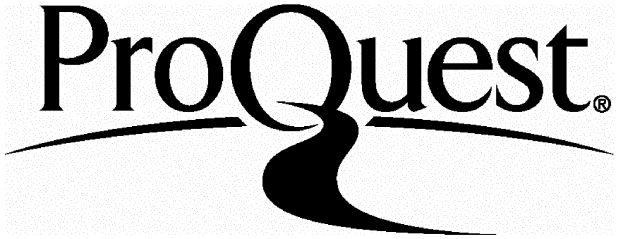
Award date:
1982

Awarding institution:
University of Bath

HYBRID-TIME ENCODED SPEECH

Submitted by Amarjit Singh
for the degree of PhD
of the University of Bath
1982

A. Singh.....



SUMMARY

Time encoded speech (TES) proposes the transmission of speech by segmenting the waveform between real-zero crossings.

Speech comprises two types of sound - voiced and unvoiced. The segment rate is low for voiced sounds and high for unvoiced sounds. Matching of the variable segment generation rate to the constant transmission rate is achieved by inserting storage buffers.

Since the variation in generation rate can be large for a particularly fricative utterance, the storage buffer has to be large, and consequently, the delay in transmission is large.

This thesis presents a technique which reduces the buffer size requirements and hence the delay in transmission. The technique, hybrid-TES, achieves these reductions by identifying the high segment generation regions of the speech waveform;and by storing and transmitting special symbols to indicate the reconstruction of these regions by spectrally shaped random noise.

LIST OF CONTENTS

LIST OF SYMBOLS

Acc	Accumulator
B	Bandwidth
BHT	Hybrid-TES reconstruction using bandpass filtered noise
β	System delay
dB	Decibel
DC,d.c.	Direct Current
DM	Delta Modulation
DPCM	Differential Pulse Code Modulation
DUR	Epoch duration boundary between voiced and unvoiced
ΔM	Delta modulation
Δ	Receiver delay
δ	Transmitter delay
EA	Epoch Amplitude
EC	Epoch Class
ED	Epoch Duration
Hz	Hertz
IPS	Intermediate Processing Stage
IEEE	Institute of Electrical and Electronic Engineers
k, prefix	Kilo or 10^3
K1 K2 K3 K4	Constants used to determine unvoiced symbol
LPC	Linear Predictive Coding
LIM	Epoch count to distinguish between voiced and unvoiced regions
M, prefix	Mega or 10^6
m, prefix	Milli or 10^{-3}
μ, prefix	Micro or 10^{-6}
N	Total number of symbols in the transmit buffer
NOE	Number of epochs
No	Number
NP	No preference

PAV	Average amplitude
PCM	Pulse Code Modulation
R	Rate of symbol generation
RMS	Root mean square
RST	Time encoded speech using repeat strategy
RZPDS	Real-zero probability distribution of speech
S, Sec	Second
Si	Number of injected symbols
SNR	Signal to noise ratio
S, prefix	Relating to silence
T	Threshold
TES	Time encoded speech
Ti	Microprocessor inspection period
Tr	Transmission rate
U, prefix	Relating to unvoiced
UV	Unvoiced
V	Voiced
>	Greater than
<	Less than
≅	Approximately
#	Immediate addressing mode
≠	Not equal to
$	Hexadecimal notation
%	Percentage
π	Pi
Ω	Ohms
t	time
MSB	Most Significant Bit
CCT	Circuit

CHAPTER 1

INTRODUCTION AND REVIEW OF SPEECH COMMUNICATION SYSTEMS

1.1 Introduction

The need for communicating thoughts and ideas between people has been recognised since the dawn of civilisation and numerous methods have been devised for this purpose.

In the early days, these ranged from speech, gestures and graphical symbols, for close communication; to smoke signals, light beams, carrier pigeons and letters transported by a variety of means, for long distance communication. With time, faster and more accurate means of long distance communication became more compelling and consequently resulted in the development of communication by electrical methods. Starting with the telegraph line, it was shortly followed by the telephone and radio transmission. Significant developments since then include radar and microwave systems, transistor and miniaturised integrated circuits, communication satellites and lasers. Today electrical communication systems span the entire world carrying voice, text, pictures and a variety of other information[1].

Irrespective of the nature of information and the actual method of transmission, a general model can be

used to describe a communication system, as shown in the block diagram of Figure 1.1. Basically, the function of the system is to transfer information from the source to the destination.

In general, the message signal produced by the source is not electrical and an input transducer is necessary to convert the message to its time varying electrical equivalent. Similarly, at the destination, an output transducer converts the electrical waveform to the appropriate message.

The source and destination are usually separated by a communication channel, which can take one of a number of forms; eg a microwave radio link over free space, a pair of wires or an optical fibre.

The transmitter couples the electrical message signal to the channel. Although it may be possible to couple the input transducer directly to the channel, it is often necessary to modify the input electrical signal for efficient transmission over the channel. Signal processing operations performed by the transmitter include amplification, filtering, encoding and modulation.

The main function of the receiver is to extract the input message signal from the degraded version of the

transmitted signal coming from the channel. It does this through demodulation and decoding, the inverse of the transmitter processes.

Due to physical limitations, communications channels have only finite bandwidth (BHz) and the information bearing signal often suffers amplitude and phase distortion as it travels over the channel. In addition to the distortion, the signal power also decreases due to the attenuation of the channel. Furthermore, the signal is corrupted by unwanted, unpredictable, electrical signals referred to as noise. Although the effects of noise cannot be completely removed, some of the degrading effects of the channel can be removed or compensated for by efficient transmitter and receiver design.

The particular requirements of the transmitter and receiver depend on the type of message to be conveyed, eg music, speech, television or data. Since this thesis concentrates on speech, an introduction is provided below to the various ways of encoding and transmitting the speech signal more efficiently.

1.2 Speech Communication

Speech communication is made possible by a complicated and interrelated pair of organs: the mouth and vocal tract, which act in combination as a transmitting apparatus; and the ear, which acts as a receiving apparatus. A speech communication system is merely an extension in space of the distance between the mouth and the ear.

As an insight into the requirements of a speech communication system it is necessary to know the nature of speech and the parameters which need to be transmitted.

1.2.1 The nature of speech[2]

Figure 1.2 shows in diagrammatic form the essential parts of the human vocal system. The vocal organs are the lungs, the trachea or windpipe, the larynx, the pharynx or throat, the nose and the mouth. The part of this 'tube' extending from the larynx to the lips is known as the vocal tract. The shape of this tract is varied extensively during speech production by moving the lips, the tongue and the jaw, ie the articulatory organs.

Speech is produced by two basic types of sound source; the class of sounds known as "voiced" sounds are produced by puffs of air released during vocal cord

vibration. Voiced sounds include all the vowels and such consonants as m, n, l and r. The other class of sounds, "unvoiced" sounds, are produced when turbulence is caused by air being forced through a narrow constriction somewhere in the vocal tract; examples of unvoiced sounds are f, s, p and t.

Some sounds such as z and v require the use of both types of sound source simultaneously.

1.2.1(a) Voiced sounds

During ordinary breathing the vocal cords are in a relaxed condition and are held fairly wide apart, but during voiced sounds they are drawn together. These cords are in fact folds of ligament at the top of the trachea, and the slit-like orifice between them is called the glottis. During production of a voiced sound air travels from the lungs up to the trachea and builds up a pressure behind the vocal cords; these are pushed apart and air rushes through the narrow glottal opening, slowing down again when it reaches the wider pharynx above.

By a combination of muscular tension in the cords and the lowering of pressure in the glottis due to the Bernoulli effect, the vocal cords are drawn back to their starting position and the air flow ceases. The

sub-glottal pressure then forces the cords apart again and the whole cycle is repeated. The vocal cords therefore act as an intermittent barrier to the flow of air from the lungs, and in fact chop the air stream so that a discrete set of puffs is produced.

The vocal cord vibration period is a function of vocal cord mass, tension and sub-glottal pressure. For normal male talkers these puffs of air are produced with a frequency typically in the range 50-250Hz, but extends to 500Hz and higher for women and children. Although the frequency of vocal cord vibration is fairly high, its rate can be changed only slowly, by varying the sub-glottal pressure and tension of the vocal cords, both of which are under muscular control. These puffs of air constitute the basic generator for the voiced sounds of speech. Typically the shape of a glottal puff (ie the volume velocity of air plotted against time) is approximately triangular and, since these puffs are quasi-periodic, can be considered to have an approximation to a line spectrum.

The vocal tract can be shown to be quite analogous (up to fairly high audio frequencies) to a mismatched non-uniform transmission line; it is therefore a resonant system that intensifies the energy of certain bands of frequencies. These resonances, whose frequencies

can be changed by movement of the articulators, are given the name "formants". The formants superimpose their response on the vocal cord signal to produce the voiced sounds of speech. Voiced sounds are usually characterised by three or four formants in frequency range up to about 4000Hz.

1.2.1(b) Unvoiced sounds

Many of the sounds of speech come into the class of sounds known as unvoiced. During unvoiced sounds the vocal cords are held wide apart and the air stream from the lungs is forced through the constriction between the tongue and the teeth as in 's', or between the teeth and the lips as in 'f', causing turbulence and producing the characteristic 'fricative' sounds of speech. The basic generator for this type of sound is the air stream whose source can be considered to be at the point of constriction. The unvoiced sounds do not exhibit the harmonic structure of the voiced sounds and the sound generator is probably best thought of in electrical terms as a random noise source. The energy of the fricative sounds is generally much lower than that of voiced sounds and the resonances in the system have greater bandwidth than for the voiced sounds.

1.2.1(c) Plosive consonants

Another class of sounds that is produced fairly often in speech are the plosives or stop consonants; these sounds are produced by stopping the flow of air from the lungs by blocking the vocal tract at some point and then very quickly releasing the air pressure. The plosives therefore are always characterised by a silence preceding the burst of energy; they may be voiced, for example, as in 'b' and 'd', or unvoiced such as 'p' and 't'.

1.2.2 Compression schemes

The frequency range occupied by the sounds of speech extends approximately from 60Hz to 10kHz and the transmission of electrical signals which are an exact analogue of the speech signal, therefore, requires a bandwidth of some 10kHz. However, transmission of a single conversation over a communications channel with a bandwidth capability of BHz, where B >> 10kHz, is wasteful in channel capacity. Multiplexing[3], where many conversations are transmitted simultaneously, enables a more useful employment of the channel. The number of conversations can be further increased by reducing the channel capacity occupied by each conversation. A tremendous amount of speech communication research effort has been and is presently directed towards this task of finding ways of reducing the

channel capacity required by each conversation.

Channel capacity is usually measured in bits per second. The theoretical maximum capacity, C, of a communication channel of bandwidth BHz for a signal to noise power ratio P/N has been shown by Shannon[4] to be

$$C = B.\text{Log}_2 \left(1 + \frac{P}{N} \right) \text{ bits/second} \qquad \ldots (1.1)$$

where N is the power of white thermal noise. For a given signal to noise power ratio the theoretical channel capacity and bandwidth are thus directly related, and methods of reducing the required capacity are often referred to as systems for bandwidth economy or compression.

These may be divided broadly into two categories. In the first the electrical signal transmitted remains an analogue of the speech sounds, even though it may be a very imperfect or distorted analogue. The methods described in sections 1.2.2(a) to 1.2.2(e) are of this type. The second category uses analysis - synthesis techniques. Narrow-bandwidth code signals are derived from the speech by means of an analyser and transmitted over the channel to the receiver. Here the code signals are arranged to control an artificial talking device, or speech synthesizer.

1.2.2 (a) Band-pass filtering[5]

If a speech signal is confined by means of a filter to the range 150Hz to 4500Hz, the intelligibility and the quality (or naturalness) are found to be quite satisfactory. Intelligibility remains adequate, though quality suffers, if the frequency band is reduced to 300-3400Hz,as in the normal telephone system.

1.2.2 (b) Frequency division[6,7]

Several schemes have been devised in which the frequencies present in the speech signal are divided by a factor of, say, two or three before transmission and multiplied back to their correct values at the receiver. An elementary way of achieving this is to record the speech and to replay it at a slower speed for transmission over a narrow bandwidth channel. Upon reception the slowed down speech is re-recorded and then played back at a faster rate to give normal speech. Although a simple arrangement of this kind enables a narrow bandwidth to be used, the actual saving of channel capacity is nil, since a greater time is needed to pass the signal. Hence, more complex processing is needed. For example, since the speech signal usually includes repetitive waveforms, it may be sampled at suitable intervals and the portions of the signal between the samples may be rejected. The frequencies contained in the sample may then be divided down and transmitted in

the same time as the original signal would have taken. At the receiving end the frequencies in the samples are restored to normal and the reduction in time which then occurs is compensated for by an appropriate number of repetitions of each sample. By use of 'Doppler' techniques it is possible to arrange for the sampling and frequency division to be carried out as a continuous process at the transmitter, and for the frequency expansion and repetition of samples to be carried out continuously at the receiver.

Experiments carried out by Gabor[6] showed that compression to one-half, followed by re-expansion, gives full intelligibility but some roughness. Speech compressed to one-quarter and even to one-sixth remains intelligible, but becomes monotonous, as the inflections of pitch are not transmitted. The intelligibility is lost only at about eightfold compression.

The concept of compression and re-expansion is easy enough to grasp, but the economies achieved are not worthy of adoption for practical systems.

1.2.2 (c) Time-assignment speech interpolation[8]

It is sometimes possible to obtain better utilisation of communication channels without actually reducing the bandwidth of the individual signals. In time-assignment

speech interpolation, for instance, the pauses in speech (which on average take up more than 50% of the time of a conversation) are filled by interleaving utterances due to different speakers. Thus, any talker may be disconnected from his channel when he stops speaking for any reason (such as to listen to a reply). The channel is switched automatically to another speaker, and when the original talker recommences to speak his conversation may be carried by quite another channel. Hence, the total number of conversations carried at peak periods can be doubled, with no loss of intelligibility or quality. The complex electronic switching apparatus is, however, very expensive and the method is only economically suitable for special circuits such as transoceanic cables.

1.2.2 (d) Interrupted speech[9]

Another way of improving the utilisation of a channel, though with some lowering of intelligibility, would be to interrupt the speech continuously at a frequency of, say, 40Hz. With equal 'on' and 'off' periods it would be possible to time-multiplex two conversations onto a single channel (see section 1.2.3).

1.2.2 (e) Speech clipping[10,11]

Under conditions of low signal to noise ratio some improvement in intelligibility can be obtained by speech

clipping. Speech has a dynamic intensity range of about 30dB and an amplifier, which is adjusted so as not to overload on loud passages, may not give sufficient amplification to the weaker sounds. By using a system of amplitude compression the intensity of weaker sounds can be raised and the intelligibility improved. The compression process may be carried to the limit giving an infinitely clipped rectangular wave of constant amplitude. This retains only the information contained in the zero-crossings of the original speech wave, yet the intelligibility is not much less than that of normal speech. Unfortunately the quality is harsh and unpleasant.

The increase in effective signal power obtained by speech clipping may approach 12dB, and from equation (1) a corresponding improvement in the capacity of the channel concerned is possible.

1.2.2 (f) Analysis-synthesis methods: The vocoder[12-14]

One way of improving channel capacity is to transmit, instead of the normal speech currents, only sufficient coded information to permit the 'remaking' of the speech by suitable apparatus at the receiving end. The coded information can be passed in the form of signals requiring a channel capacity less than that of the speech currents, so that the line or other transmission system is able to

accommodate more channels within its pass range of frequencies. This is the basis of speech analysis-synthesis telephony systems, such as fixed channel vocoders and parametric (formant tracking) devices.

The formant tracking vocoder attempts to measure in real time the frequency and amplitude of spectral peaks of the speech signal (the formants) and transmits these measurements as parameters to control a synthesiser as shown in Figure 1.3. At the same time the larynx vibration rate and the decision as to whether the speech is voiced or unvoiced must also be measured and transmitted as additional control parameters. It is generally recognised that good speech quality can be obtained from a formant synthesiser, but when connected to a real-time analyser the results obtained to date have always been inferior. This is mainly because of the difficulty in measuring formant frequencies accurately. Figure 1.3 shows that the formant frequency ranges overlap and this tends even more to aggravate the problem. The advantage of the formant vocoder over most other types of vocoder is the bandwidth compression that can be obtained. Typically a formant vocoder needs eight parameters, and each of these can be bandlimited to about 25Hz, so that a bandwidth reduction of the order of 20:1 can be obtained relative to the original speech band. It should, however, be emphasised that all the difficulties of real-time

formant analysis are not yet solved, and until they are it is unlikely that formant vocoders will prove to be very useful.

The fixed-channel or 'frequency-band' vocoder was first demonstrated by Dudley and has undergone extensive development. In this system an analysis of the power spectrum of speech is made by means of a number of band-pass filters, which divide the audio-frequency spectrum into, say, twelve adjacent bands (Figure 1.4).

The energy in each band is measured, and twelve narrow bandwidth signals are obtained, which vary as the energy varies. An additional signal gives the pitch of the larynx signal, and is also used to specify whether the speech is voiced or unvoiced. At the receiving end this pitch signal is used to switch into circuit, either a relaxation oscillator which acts as a 'buzz' source for voiced sounds, or a noise generator which acts as a 'hiss' source for unvoiced sounds (Figure 1.5). The pitch signal also controls the fundamental frequency of the buzz source.

The line spectrum of the buzz source, or the continuous spectrum of the hiss source, is filtered into twelve frequency bands, corresponding to those of the band-pass filters of the analyser. The magnitude of the

output in each band is controlled by the appropriate line signal, using a system of modulators. The output of these modulators are then combined to produce artificial speech similar to the original. An overall bandwidth compression of 10:1 in the transmitted signal is possible.

Most modern vocoders have been designed to use a digital transmission path. There are two basic reasons for this. First, through the use of small general purpose digital computers speech researchers have been able to apply a wide variety of digital signal processing techniques to speech communication problems. These techniques cover a range of complexity and sophistication that is impossible to match with analogue methods. Second, the recent and predicted future developments in integrated circuit technology make it possible to realise digital speech processing schemes economically as hardware devices having the same sophistication and flexibility as a computer program implementation.

To digitise a channel vocoder both spectrum and pitch channels have first to be multiplexed and then coded by some digital coding scheme, such as pulse code modulation or deltamodulation. At the receiver a digital-to-analogue converter reconstitutes the signals into an analogue form; these analogue signals are then

filtered by low-pass filters with cut-off frequencies of the order of 25Hz, and are then used as control signals to the synthesiser. Digit rates used are in the order of 2000-3000 bits per second; which can be transmitted by modern data-modems over normal 3kHz lines.

1.2.3 Digital representation of speech signals

Conceptually, the simplest digital representations of speech are concerned with direct representations of the speech waveform. Such schemes are based on Shannon's sampling theorem, which states that any bandlimited signal can be exactly reconstructed from samples taken periodically in time if the sampling rate is twice the highest frequency of the signal.

1.2.3 (a) Pulse Code Modulation (PCM)[15,16]

Pulse code modulation is a waveform encoding technique and can be used for converting any analogue signal into digital form, for use by computers and other digital equipment, for subsequent processing.

It comprises three basic processes: sampling, quantising and encoding. The amplitude range of the sampled data is divided into a finite number of discrete levels; a given sample amplitude is then referred to the nearest level and a digital code is generated. Figure 1.6 illustrates the quantisation operation. The difference

between the original signal and the quantised signal is regarded as noise and is called 'quantisation noise'. It has been shown[16] that the quantisation noise, Q_N, is a function of the step size, S, where

$$Q_N = \frac{S^2}{12} \qquad \ldots (1.2)$$

Thus, if the steps are uniform in size, small-amplitude signals have a poorer signal-to-quantisation noise ratio than large-amplitude signals. To correct this situation, within the constraints of a fixed number of levels, it is advantageous to taper the step size so that the steps are close together at low signal amplitudes and further apart at large amplitudes. Such variation of step size yields a signal-to-noise ratio improvement for small signals, although strong signals are somewhat impaired. It has been determined[16] experimentally that, rather typically, the instantaneous speech signal amplitudes are less than 25% of the rms signal value for 50% of the time. Consequently, the tapering of steps is very useful in connection with speech signals.

While it is possible to build a quantiser with tapered steps, it is more feasible to achieve an equivalent effect by distorting the signal before application to the quantiser. It can be achieved by a non-linear network having the input-output characteristic as shown in Figure 1.7. Such a network is called a compressor.

An inverse distortion at the receiver ensures overall distortionless transmission and the complete transmitter and receiver operations of compressing and expanding are often referred to as companding.

Practical telephony systems using PCM are designed for eight bit per sample representation with a sampling rate of 8kHz, thus resulting in a transmission rate of 64kbits/second.

1.2.3 (b) Delta Modulation (DM) [17]

Delta modulation is another technique for converting analogue signals into digital form. It differs from PCM in that it transmits information about the derivative of the input signal rather than its instantaneous amplitude. The basic scheme is illustrated in Figure 1.8.

The analogue input signal is encoded by the delta modulator into binary pulses which are conveyed to the terminal equipment for transmission. These pulses are also locally decoded back into an analogue waveform, by an integrator in the feedback loop, and subtracted from the input signal to form an error which is quantised to one of two possible levels, depending on its polarity. The quantised output is periodically sampled to produce the output binary pulses. In this way the decoded waveform is made to 'track' the input signal in small steps.

This process of waveform following in small steps makes deltamodulation particularly suitable for signals in which differences between successive ordinates are small. Thus, the low frequency dominance in the speech signal can be directly accommodated. However, the problem arises when the step size is of a fixed value. During any period of time when the changes in the input signal are less than the step size, the system no longer follows the signal, and a train of alternating positive and negative pulses is produced. Similarly the system overloads when the slope of the signal is too high (see Figure 1.9). Adaptive deltamodulation allows both of these complaints to be relieved by varying the step size in accordance with the signal being encountered.

Experiment[16] has shown that, for good quality speech transmission, DM, in its simplest form, requires more bandwidth than PCM, where the relative transmission rates are 64 kbits/sec and 100 kbits/sec respectively. On the other hand, DM has the advantage that the hardware required for its physical implementation is very much simpler than that required for PCM. Combining the simplicity of DM with the economy of PCM results in differential PCM (or DPCM)[18] and gives good quality speech for a much reduced transmission rate (32 kbits/sec). In this case the difference signal available from the comparator is applied to the input terminal of the PCM

system. Lower transmission rates are possible with use of level adaptation, but at a loss in the quality of transmission.

1.2.3 (c) Predictive coding[19,20]

In this coding method both the transmitter and the receiver estimate the signal's current value by linear prediction on the previously transmitted signal. The difference between this estimate and the true value of the signal is quantised, coded and transmitted to the receiver. At the receiver the decoded difference signal is added to the predicted signal to produce the input speech signal.

A block diagram illustrating the principle of predictive coding is shown in Figure 1.10. The input signal s(t) is sampled at the Nyquist rate to produce the samples s_n of the signal. The predictor forms an estimate $\hat{s}_n$ of the signal's present value based on the past samples, r_{n-1}, r_{n-2}, ... , of the reconstructed signal at the transmitter. The predicted value $\hat{s}_n$ of the signal is next subtracted from the signal value s_n to form the difference δ_n, which is quantised, encoded and transmitted to the receiver. At the same time, the transmitted signal is decoded at the transmitter and the signal reconstructed in exactly the same manner as is done at the receiver. The reconstructed signal is then used to

predict the next sample of the input signal.

At the receiver the transmitted signal is decoded and added to the predicted value of the signal to form the samples r_n' of the reconstructed signal. The predictor used at the receiver is identical to the one employed at the transmitter. The samples r_n' of the reconstructed signal are finally low-pass filtered to produce the output signal r'(t).

The exact form of the predictor for the speech wave will not be described here, since it is not an essential part of this survey; but it will suffice to say that it depends on the model used to represent the human speech production process and that ample literature is available on the subject.

The predictive coding system was simulated on a digital computer by Atal and Schroeder[20], and they showed that it is a promising approach to digital encoding of speech signals for high-quality transmission at substantial reductions in bit-rate. Unlike past speech coding methods based on the vocoder principle, the prediction coding scheme attempts to reproduce accurately the speech waveform rather than its spectrum. Subjective tests showed that there was only slight, often imperceptible, degradation in the quality of the reproduced speech.

Their studies pointed out that the binary difference signal and the predictor parameters together can be transmitted at bit rates of less than 10 kbits per second, or several times less than the bit rate required for PCM encoding with comparable speech quality.

1.2.3 (d) Time encoded speech (TES)[21-24]

A more recent addition to the above techniques of speech processing for capacity reduction is time encoded speech. This method of low-bit rate encoding has been found by considering the 'perception' rather than the 'generation' of speech; as, for example, in the case of vocoders, which are modelled entirely on consideration of the human vocal tract. The aim of TES is simply to present to the ear acoustic cues which produce the correct physiological responses.

The crudest form of TES and one long known, is infinitely clipped speech, which retains only the zero crossing information of the original speech waveform, yet maintains a high level of intelligibility and some speaker recognition. The central concept of this approach is that instead of encoding instantaneous values of the speech waveform at regular intervals, as with PCM, changes of value, as with DM, or descriptions of the speech spectrum, as in channel and formant tracking vocoders, time encoding depends on the transmission of

coded shape descriptions for successive extended segments of the speech waveform. It is broken into segments between successive real zeros of the waveform. For each such segment the code consists of a single digital word. This word is derived from two parameters of the segment: its quantised time duration and its shape. The measure of duration is straightforward and uses logarithmic quantisation so that short intervals are transmitted more accurately, thus maintaining an approximately constant fractional accuracy.

The shape of the segment is compared with a 'catalogue' of shapes, and a code is selected which identifies the shape in the catalogue nearest to the actual segment shape. A complication of this code selection is that shapes which look different may not necessarily sound different, since phase relationships that greatly affect waveshape are hardly perceived by the ear. Research is still being conducted on attempting to classify shapes as distinct only when they sound different.

Reconstruction of TES is a relatively simple matter, in that, stored segment forms are reproduced in sequence at the correct duration, in accordance with the shape and duration specifications corresponding to the received symbol.

Due to the nature of speech, this technique of encoding suffers from the problem of the high segment generation rate and hence the requirement of a high transmission rate, during unvoiced sounds. As a result, much research is being conducted to optimise the technique and thus achieve the lowest bit rate of transmission.

In summary, Table 1.1 shows the relative advantages and disadvantages of some of the more popular encoding schemes.

1.3 Organisation of the Thesis

This thesis looks into the problem, for TES, of the variable segment generation rate and proposes an alternative encoding strategy which takes account of this variation. It is broadly divided into three sections. The first section (Chapters 1 and 2) comprises the introductory comments and the proposal of hybrid-TES encoding, which is based on the conclusions derived from the study of the real-zero probability distribution of speech.

The second section (Chapters 3, 4 and 5) presents the feasibility study of the encoding scheme, where: Chapter 3 details the identification of the high epoch generation regions (unvoiced sounds); Chapter 4 shows the investigation of the reconstruction parameters;

Class	Examples	Bit rate kbits/sec	Quality	Complexity	Current status
Waveform	PCM	64	Excellent	Low	Widespread use
	Δ modulation (DPCM)	32	Very good		
		16	Fairly good		
	(with companding & level adaption)	8	Poor		
Analysis/ synthesis	LPC	2.4	Fair	High	Some use, increasing
	Channel vocoder	2.4	Fair	High	Widespread use
	Formant vocoder	1.2	Can be good	Very high	Research
TES	-	4-16	Fair	Medium	Research

TABLE 1.1 Comparison of different encoding systems (from Holmes, private communication)

Chapter 5 analyses the performance of the hybrid-TES encoding with respect to the buffering requirements and the quality of reconstruction.

The third section (Chapters 6 and 7) provides the summary and conclusions on the performance of the encoding scheme, where: Chapter 6 makes a formal presentation of the hybrid-TES encoding technique, bearing in mind the real-time considerations; and Chapter 7 provides the conclusions to the thesis.

CHAPTER 2

THE REAL-ZERO PROBABILITY DISTRIBUTION OF SPEECH (RZPDS)

2.1 Introduction

Time encoded speech is a recent proposal, and as such the problem still remains of optimising the technique to achieve the lowest bit rate of transmission. In order to associate an optimum number of bits with the representation of real-zero intervals of speech it is necessary to have some knowledge of the typical occurrence of these intervals. This information can be provided by means of a long term probability against real-zero interval distribution.

A real-time method for measuring the long-term probability distribution of speech has been suggested by Davenport[25], but it is very slow and cumbersome. A new method, which is much faster and takes account of the present day technology, makes use of a microprocessor and is presented in this chapter.

Measurement of the long-term probability distribution through non-real-time processing is not practicable due to the large storage space necessary for the utterances. However, since the subsequent work is to

be conducted in simulation short-term measurements are performed on the PDP-11 minicomputer to confirm that similar effects, to the real-time measurements, are obtained with sampled signals.

The results of measurement have shown that the real-zero probability distribution of speech comprises two peaky regions: one in the 1-2ms range, due to the voiced sounds; and the other, in the range 0-0.5 ms and sometimes much larger, is primarily due to the unvoiced sounds. In particular, for speakers with highly pronounced fricatives, the voiced peak is often "swamped" by the unvoiced peak and a clear distinction cannot be made.

A further insight into the pattern of the probability distribution is provided by measurement of the epoch rate against time, which shows that the short duration epochs, in the unvoiced peak, occur in bunches. It is this latter result which provides impetus for the proposal of an alternative coding strategy for TES, and hence a saving in the transmission rate requirement.

2.2 Real-Time Measurement of RZPDS

Defining To(t) as the duration of the zero-crossing period existing at the time t, the probability $p(T_1)$ defined by the equation:

$$p(T_1) = p[T_1 < To(t) < T_1 + dT_1] \quad \ldots (2.1)$$

is the probability that at a given instant of time T_1 the duration of the corresponding zero-crossing period falls in the interval $(T_1, T_1 + dT_1)$. The first order probability density function $w_1(To)$ is defined by the equation:

$$w_1(T_1)\, dT_1 = p(T_1) \quad \ldots (2.2)$$

2.2.1 Davenport's method

The problem of measuring the zero-crossing probability distribution concerning speech waves was first tackled in 1952 by Davenport[25]. Data was collected by using the equipment shown in the block diagram of Figure 2.1.

The speech wave x(t) was clipped to form the zero-crossing wave z(t), which was then applied to the period-pulse generator to generate an output pulse whenever z(t) changed value. The amplitude of this output pulse was proportional to the zero-crossing period. The level selector was so adjusted that it generated a pulse whenever the amplitude of its input pulse corresponded to a zero-crossing period whose duration was in the interval $(T_1, T_1 + dT_1)$. Counter No 1 made a count of the total number 'n' of period pulses, and Counter No 2 counted the number 'V' of

period pulses corresponding to the event $|T_1 < To(t) < T_1 + dT_1|$. It was shown that:

$$p(T_1) \cong V.T_1/T \qquad \ldots (2.3)$$

and as a result

$$w_1(T_1) \cong V.T_1/T.\Delta T_1 \qquad \ldots (2.4)$$

where T is the total duration of the observation interval. These expressions become exact only in the limits, $T \to \infty$ and $\Delta T_1 \to 0$.

At attempting to measure the first probability distribution Davenport immediately came across the problem of the unavoidable presence of system noise, which is characteristically similar to unvoiced speech and has the effect of increasing the zero-crossing rate during the inter-word and inter-sentence passages. This in turn leads to the question: How much of the combined distribution is due to the speech wave alone? A solution to the question was obtained on the assumption that if the unvoiced sounds had significantly larger amplitudes than the amplitude of the system noise, it should be possible to separate the two effectively. The separation was accomplished by the addition of a low-frequency periodic bias signal having an amplitude larger than that of noise, but smaller than that of the unvoiced sounds.

Davenport concluded that the results hence obtained showed fairly good correspondence to his theoretical predictions, which were based on the measurements of the first probability distribution $w_1(To)s + n$ for a speech wave and system noise, of $w_1(To)n$ for system noise alone, and the assumption that the unvoiced sounds are significantly larger in amplitude than the system noise. As a result of the assumption it was possible to say (approximately) that at any given time either the speech wave or the noise is present in the system, and that these are mutually exclusive events whose probabilities are additive. Using the definitions $p(T_1)s+n$, $p(T_1)n$ and $p(T_1)s$ as the probabilities that the combined speech and noise wave, the noise wave alone and the speech wave alone respectively, have zero-crossing periods of duration $|T_1 < To < T_1 + dT_1|$ it was shown that:

$$p(T_1)s+n = D.p(T_1)n + (1-D).p(T_1)s \quad \ldots (2.5)$$

where D is the fractional dead time (ie the fraction of time in which there is no speech in the combined wave). A value of 0.22 was used for D to solve equation (2.5) for the probability distribution of speech wave alone, $p(T_1)s$.

2.2.2 The microprocessor method[26]

Davenport's method was very cumbersome and time-consuming because the level selector had to be readjusted

to a different level after every set of measurements and the measurements had to be repeated for the new level. An alternative form of measurement, which is faster and more efficient, has been designed by the author and makes use of a microprocessor to perform the timing and counting functions. It uses the equipment shown in the block diagram of Figure 2.2.

The zero-crossing detector compares the speech waveform with zero and results in a rectangular wave output with the level depending on the polarity of the input. A typical speech wave input to the detector and its output are shown in Figure 2.3. Its circuit is given in Figure 2.4.

The rectangular wave output from the detector is applied to the microprocessor input, where the change of sign is recognised as an indication of the zero-crossing and a flag is set. The timing and counting routine regularly inspects the flag and takes the appropriate action, which is basically to measure the time between zero-crossings and increment the corresponding counter. Although details of the microprocessor software design are given in Appendix A, the flow diagram for the timing and counting routine is shown in Figure 2.5

Depending on the conditions the algorithm may be in any of the three loops. Thus, to ensure that the flag inspection is made at a regular interval, the sum of instruction times around any loop must always equal the sampling rate (50μs). This requires that the time delays are of different value - hence the apparent duplication of part of the program code.

Each interval duration is assigned a counter and the program terminates when any one of the counters overflows. The numbers stored in the counters are then output, by means of an X-Y plotter, in the form of a graph; where the vertical axis represents the number of times a particular interval duration occurs and the horizontal axis, the different interval durations (see Figure 2.6).

As with Davenport's method, the noise problem still has to be resolved. A means of overcoming this, to a certain extent, is available in the use of a 'threshold' voltage. It has a similar effect to the bias voltage introduced by Davenport, but is easier to implement. In practice it means that the zero reference level at the comparator is made into a variable pre-set reference level. The value at which it is set has to be determined by caution: for too low a value allows the noise voltages to switch the output waveform, and too high a

value causes the small amplitude speech variations to be neglected. Thus, a compromise has to be made.

In practice, however, it is not a very difficult decision, since when the reference voltage is set to too low a value the output of the comparator, with no speech input, it undefined due to the random variations in the system noise voltage. The reference voltage is then increased until the comparator output becomes defined. The minimum value at which this happens is called the "threshold" voltage.

2.2.3 Result analysis

Measurements were taken to show:

(i) The effect of the threshold voltage on the RZPDS.

(ii) The effect of filtering on the RZPDS.

(iii) The individual voiced and unvoiced distributions.

(iv) The long-term pattern of the complete distribution and the effect on it of changing the language of speech.

The conditions and settings under which these measurements were performed are listed in Table 2.1.

The results are presented graphically and are best analysed by investigating, one at a time, each of the cases mentioned above. It should be noted that the vertical axes on the graphs have been scaled only to

facilitate comparison between the different regions. They are not meant for numerical comparison between the different graphs unless they are obtained from measurements made under identical conditions and settings. More importance is placed on recognising the shape characteristics than on actual numerical values.

2.2.3(a) The threshold voltage

Figure 2.7 shows the probability distributions of a speech recording; taken firstly with a 20Hz square wave signal of amplitude 100mV (as suggested by Davenport); and secondly with a d.c. threshold voltage of value 100mV. The distributions tend to follow very similar patterns, but a closer resemblance is shown by reducing the number of class intervals. Using Sturges[27] rule, which states that given a number of samples 'x' the optimum number 'N' of class intervals is given as:

$$N = 1 + 3.3 \log_{10}(x) \qquad \ldots (2.6)$$

the graphs of Figure 2.8 are obtained. A numerical comparison between graphs (a) and (b) of Figure 2.7 shows that there are differences. These differences are caused by the uncertainty of the starting point on the speech recording.

Variation of the threshold voltage shows a noticeable effect on the distribution pattern. This effect can be

explained by reference to Figure 2.9. If the amplitude of the speech waveform is small, as shown in Figure 2.9(a), variation of the threshold voltage has the effect of altering the output waveform from the comparator. However, with a large amplitude speech signal (Figure 2.9(b)) the same variation in threshold voltage has little effect on the output waveform. Thus, since recordings of speech invariably differ in amplitude (loudness) a comparison of different speech recordings cannot be made at any particular threshold voltage. Instead, a separate threshold voltage has to be located for each of the speech recordings, as the threshold only accounts for the system noise and not the acoustic noise.

Figure 2.10 shows the effect of varying the threshold voltage. As the threshold is increased the zero-crossings due to the system noise (the first 50μs interval) are reduced. An additional effect is to shift the position of the secondary peak in the distribution.

2.2.3 (b) Filtering the speech

A commercial, 8th order Barr and Stroud, filter was used to perform the band-pass filtering operation. The filter has two modes of operation - normal and damped. The normal mode provides a response similar to that of an 8-pole Butterworth function, whereas the damped mode provides an improved filter phase response, which effectively reduces ringing and overshoot on pulse and

step-type waveforms. In order to avoid the ringing effect interfering with the zero-crossing measurements the damped mode of operation was selected.

Measurements (Figure 2.11) were taken to show the effect on the distribution of bandlimiting speech between 300Hz-3400Hz. The higher frequency components (< 300μs) are noticeably reduced, although not completely eliminated. This is because of the gradual nature of the cut-off frequency. The low frequency components (> 1.5ms) remain relatively unaffected. Figure 2.12 illustrates that extending the low frequency limit has little or no effect on the distribution. It was measured with the bandwidth set to 100Hz-5000Hz.

2.2.3 (c) The voiced/unvoiced distributions

The real-zero probability distributions for voiced sounds and unvoiced sounds were determined individually, to verify the voiced/unvoiced nature of the complete long-term distribution.

A recording was made up of voiced sounds which made use of the vowels and such consonants as 'm', 'n', 'r' and 'l'. To avoid the infiltration of any unvoiced sounds, the pitch of the sounds was varied slowly and with care. The complete recorded passage lasts approximately one minute.

The results of performing the measurements on the recording are presented in Figure 2.13 and they show a prominent peak in the range between 1ms and 2.5ms, which corresponds with a frequency range between 200Hz and 500Hz.

The unvoiced speech recording was constructed from a combination of such utterances as 'sh', 's', 'f', 't' and 'p' and it lasts approximately one minute. The resulting distribution is presented in Figure 2.14 and shows a peak in the range 100μs to 500μs. This time interval corresponds with a frequency interval between 1kHz and 5kHz, which is the usual range of frequencies for unvoiced sounds.

2.2.3 (d) The complete distribution

Two recordings of speech were made by the same speaker, to note the effect on the distribution of changing the language of speech. One recording was of speech in the English language, obtained by reading a passage from a technical paper lasting approximately two minutes. The other recording was of speech in the Punjabi language, and was obtained by reciting a number of short sentences, lasting a total of about two minutes. Measurements were performed on both of these recordings separately and they both gave similar results.

Figure 2.15 shows the distribution for the English language. There is a peak between 1ms and 2.5ms, which corresponds with the peak due to the voiced sounds in Figure 2.13. A second peak in the range 100μs to 600μs corresponds to the unvoiced sounds of Figure 2.14. The distribution for the Punjabi language is shown in Figure 2.16 and it further illustrates the voiced/unvoiced nature of the distribution. Comparison between Figure 2.15 and Figure 2.16 shows that the peak due to the voiced sounds occurs between the same time intervals for both of the languages. The high number of zero crossings in the first interval (0-50μs) is possibly due to the high level of background noise in the recordings.

Figure Number	Threshold Voltage	Bandwidth	Type of Speech
2.10(a)	0mV		Hindi
2.10(b)	40mV		Hindi
2.10(c)	80mV		Hindi
2.10(d)	120mV		Hindi
2.11(a)	60mV	300Hz-5kHz	JSRU
2.11(b)	60mV	300Hz-3.4kHz	JSRU
2.12(a)	120mV	100Hz-5kHz	JSRU
2.12(b)	120mV	300Hz-3.4kHz	JSRU
2.13	50mV	100Hz-5kHz	Voiced
2.14	50mV	100Hz-5kHz	Unvoiced
2.15	25mV	100Hz-5kHz	English
2.16	25mV	100Hz-5kHz	Punjabi

TABLE 2.1 The conditions and settings used for real-time measurements

2.2.4 Conclusions

A method of measuring the real-zero probability distribution of speech, using a microprocessor, has been devised. Its main advantage over Davenport's method is in the ease and speed with which the results can be obtained.

The threshold voltage has been shown to have the same effect on the distribution as the low frequency bias voltage used in Davenport's method, ie it reduces the effect of system noise. Its value, however, is critical and depends on the signal to noise ratio of the speech recordings.

Bandlimiting of speech causes a more noticeable effect on the higher frequencies. Reducing the higher limit of the bandwidth reduces the fricative frequencies of speech, whereas, changing the lower limit has little effect on the distribution.

The real-zero probability distribution pattern corresponds well with Davenport's predictions and comprises two peaks: one due to the voiced sounds; and the other, larger one, due to the unvoiced sounds. This has been confirmed by measuring the distributions of voiced and unvoiced sounds separately. Also, changing

the language of speech from English to Punjabi has not shown any differences in the long-term pattern of the distribution.

2.3 Non Real-Time Measurement of RZPDS

Non real-time measurements were taken on the PDP-11 computer to confirm (or otherwise) the general nature and the voiced/unvoiced distinction of the probability distribution.

Digitised speech files of several utterances of different durations, sampled at 10kHz and bandlimited between 10Hz and 4.5kHz, were used for measurement purposes. Although context of the utterances was quite different from that used for real-time measurements, the general pattern of the distribution was expected to be the same.

2.3.1 Location of the zero-crossings

Since the utterances are only defined as discrete samples of the original waveform, x(t), exact determination of all zero-crossing times is impossible. However, close approximations can be obtained by locating the regions of zero-crossings and using simple local interpolation to estimate the actual zero crossing times. Figure 2.17 shows the geometry of a typical zero-crossing. X_1 and X_2 are discrete samples at either side of the

zero-crossing and are of opposite polarity, thus indicating the occurrence of a zero-crossing.

The simplest estimation to the actual zero-crossing is to regard it as occurring at one of the sample positions, for example, the sample position after the zero-crossing, at $(Ts - \tau)$. The error in this estimation, normalised to the sample interval, Ts, is given as:

$$\text{error} = \left(1 - \frac{\tau}{Ts}\right) \quad \ldots (2.7)$$

In the case of zero sample values, the ambiguity of the zero-crossing position may be resolved by considering adjacent samples (Table 2.2).

The estimation error can be considerably reduced by interpolating between the sample values on either side of the crossing. First-order interpolation between X_1 and X_2 (Figure 2.17) yields an estimated zero-crossing at a time t_1 after X_1. The estimation error, ε, in this case is given by:

$$\text{error} = \frac{\tau}{Ts} = \frac{(Ts - \tau).X_1 + \tau.X_2}{X_1 - X_2} \cdot \frac{1}{Ts} \quad \ldots (2.8)$$

Evaluation of this expression requires knowledge of the function X(t). However, it is clearly evident from Figure 2.17 that a much closer estimate to the actual zero-crossing is obtained. Higher order interpolation would obviously give closer estimates of

zero-crossings, although it would considerably increase the computational load. An Nth order polynomial interpolation requires the solution of (N+1) simultaneous equations for determination of the polynomial coefficients in the neighbourhood of the zero-crossing, followed by a formula solution for the Nth order polynomial. Since the interval durations for TES will be encoded somewhat approximately, the increased accuracy yielded by higher order interpolation does not seem, in practice, to be necessary. Therefore location of zero-crossings, for the study of the RZPDS, is made by first-order interpolation.

2.3.2 Testing the software

The first-order interpolation strategy is incorporated into a program, EPOCH1, which is designed to measure real-zero interval (or epoch) durations in a speech file. It accepts as input the digitised speech file and outputs a results file, which contains a list of the successive epoch durations. Figure 2.18 shows its flow diagram.

The zero-crossings are detected, by monitoring the signs of successive sample values, in accordance with Table 2.2. On detection, first-order interpolation allows estimation of the zero-crossing time and hence the evaluation of XB, the fraction of the sampling interval

at which the zero-crossing is estimated. The interval duration between successive zero-crossings is determined by adding the number of complete sample intervals, NDUR, to the fractions, XB and XC, evaluated by interpolation. $XC = 1 - XB_{-1}$, where XB_{-1} is the interpolated fraction of the previous epoch. For the first epoch XC is set to zero.

The program was tested for correct operation by attempting to measure zero-crossing intervals of a known test signal. Execution of the program ASCOSS, generates a test signal file, which contains samples of the two-component signal:

$$y(t) = 100.0 \cos \left(0.11 \times \frac{2\pi}{Ts} \times t\right) + 150.0 \cos \left(0.055 \times \frac{2\pi}{Ts} \times t\right) \quad \ldots (2.9)$$

Figure 2.19 shows a sketch of the waveform, which has a fundamental frequency of 550Hz. The test file contains 4 blocks, each of 256 sample points. Since the function has a period of 18.18 sample intervals each block of the file contains 14.08 periods, or 28.16 epochs. Thus, the complete file contains 112.64 epochs. Execution of the analysis program EPOCH1 on this test file resulted in 112 whole epochs being detected. The measured values of epoch durations were:

$$1.159 \pm 0.008 \text{ ms}$$
$$0.659 \pm 0.0015 \text{ ms}$$

and the calculated values were:

1.163 ms
0.655 ms

The test thus verified the following aspects of the epoch measuring program:

(i) General measurement of all epochs in the file, under conditions of non-integer relationship between the sampling rate and the test signal.

(ii) Smooth follow-on from one block of the file to the next.

(iii) Accuracy of the first order estimation algorithm.

2.3.3 Result analysis and conclusions

Having confirmed the algorithm and the program implementation of it the available speech files (listed in Appendix B) were analysed. The results files were further processed by program HPLOT2 to arrange the data in graphical format. Execution of the available GRAFIX package (GRPLOT, GRAFIX) on the formatted files allows the data to be output via the X-Y plotter. It is noted that the GRAFIX package results in "point" plots rather than histograms, as in the previous case.

Figure 2.20 shows the probability distribution for the utterance APPLE8, and the voiced/unvoiced distinction can be clearly observed. The distributions for the

remaining utterances are summarised in Figure 2.21, where similar effects can be seen. However, in the distributions for the utterances CBONLY and BIRD the unvoiced peak is much more prominent. This result is not contradictory to the previous conclusions, but merely an extreme case of the unvoiced peak totally swamping the voiced peak. This was due to the high fricative content of the utterances, which was revealed after listening to the recordings of the utterances.

Sample	x_0	x_1	x_2	x_3	Result
	-	0	+		Zero-crossing at x_1
	-	0	-		Point of contact at x_1
	-	0	0	+	Zero-crossing at x_2

TABLE 2.2 Resolution of zero-crossing ambiguity

2.4 The Proposal of Hybrid-TES

Time encoding of speech depends on the transmission of coded shape descriptions for successive extended segments of the speech waveform. The waveform is segmented between successive real-zeros of the function. For each such segment of the waveform the code consists of a single digital word. The transmission rate therefore depends on the average occurrence of the real-zero segments or intervals of speech.

The probability distribution measurements show that the real-zero intervals for voiced sounds are generally of 1ms-2ms duration. For unvoiced sounds they are of 0-0.5ms duration, and in particular for the unvoiced fricative sounds they are in the range 50μs-0.3ms. Also, examination of the speech waveform shows that the real-zero intervals of the unvoiced sounds are generally bunched together and occur in sequence for long durations. As a result, the interval generation rate during these unvoiced sounds is very high. Indeed a plot of average interval generation rate against time for one of the more fricatively emphasised speech files confirms this.

A program AERAT1 is designed to take the results file obtained from executing EPOCH1, analyse it and present the data in a form suitable for plotting by the GRAFIX package. Figure 2.22 shows the resulting graph of the average interval generation rate (over 50 epochs) against time for the speech file CBONLY. The peaks in the graph are the result of unvoiced sounds, 'ch', 's', 't', and 'k' in the utterance "Charles bottleneck". In general, the rate of real-zero intervals varies from low (≅500/second) in the voiced region to very high (≅6000/sec) in the unvoiced fricative region.

The use of available communication channels requires that the coded information be transmitted at a constant

rate. Matching of the variable interval generation rate to the constant transmission rate is possible by inserting a storage buffer at the transmitter and at the receiver. In order to cater for the very high generation rate during the fricative sounds the transmit buffer capacity has to be large. Also, to ensure reconstruction of small duration intervals without distortion a delay must be introduced at the receiver.

It is in consideration of these requirements that an alternative encoding strategy is proposed for TES.

Since the high generation rate and small duration intervals are generally due to unvoiced sounds it is here that economy must be exercised. Many[28-31] researchers have previously concluded that unvoiced sounds are excited by a random source. Since for each of the limited number of unvoiced sounds the vocal tract performs a filtering operation, it should be possible to encode each sound by a single parameter (signifying the filter characteristics) instead of many epochs, and reconstruct it by filtered noise. In this way not only is the overall transmission rate expected to be reduced, but also the buffer length and consequent delay.

An investigation of the feasibility of this proposition is conducted and is presented in Chapters 3, 4 and 5. The following general approach is adopted:

(i) Identification of the unvoiced or high interval generation regions of the speech waveform.

(ii) Substitution of the identified unvoiced regions by bandlimited random noise to determine the types of parameters required for encoding and reconstruction.

(iii) Comparison and quantification of wholly TES reconstruction and hybrid-TES reconstruction regarding: speech quality, probability distributions, transmission rates, buffer lengths and delays.

CHAPTER 3

IDENTIFICATION OF THE UNVOICED REGIONS OF THE SPEECH WAVEFORM

3.1 Introduction

For evaluation of the hybrid-TES proposal it is necessary to initially identify the regions of the speech waveform which give rise to the high transmission rate. The identified regions can then be encoded to allow the relevant substitutions to be made during reconstruction. Since the increase in the transmission rate requirement is primarily due to the unvoiced sounds of speech, it is necessary to distinguish between these unvoiced sounds and the rest.

3.1.1 Review of previous methods

Algorithms[32-35] have been designed in the past, which identify the voiced and unvoiced regions of speech in conjunction with pitch analysis. For example, in the cepstral pitch detector[33] the voiced-unvoiced (V-UV) decision is made on the basis of the amplitude of the largest peak in the cepstrum. As pointed out by Atal and Rabiner[36] there are two disadvantages in this approach to general V-UV decision making. First, the decision is based on a single feature - the degree of voice periodicity. Voiced speech is only approximately periodic; sudden changes in articulation and the idiosyncracies of vocal

cord vibrations can produce speech waveforms which are not periodic. In such cases, a feature such as the amplitude of the largest cepstral peak will fail to distinguish voiced speech from unvoiced. In practice, additional features such as the rate of zero-crossings of the speech waveform, the ratio of low to high-frequency energy, etc, must be included in the decision procedure. Secondly, the V-UV decision is tied to the pitch detection which may be acceptable for speech synthesis applications. However, for other applications, such as speech segmentation or speech recognition, the linking of V-UV decision to pitch detection can result in unnecessary complexity as well as in poorer performance, particularly at the boundaries between voiced and unvoiced speech. For pitch detection, a large speech segment, 30-40ms long*, is necessary, which can result in unwarranted mixing of voiced and unvoiced speech. By separating the V-UV decision from pitch detection, it is possible to perform the V-UV decision on a much shorter segment, thereby enabling faster tracking of the changes from one class to another.

Consequently Atal and Rabiner[36] proposed a pattern recognition system to achieve a voiced-unvoiced-silence classification of speech on the basis of the measurement of five features from successive 10ms segments of speech. The five features measured were: energy of the signal; zero-crossing rate of the signal; autocorrelation

* More recently shorter segments have been used

coefficient at unit sample delay; first predictor coefficient; and energy of the prediction error. The method assumes a normal distribution for the variation of the features and computes a statistical distance measure from the set of feature measurements to classify the segment. Choice of the features, in this case, was based partly on the experimental evidence that the parameters vary consistently from one class to another, and partly on the knowledge of the method in which voiced and unvoiced sounds are generated in the human vocal tract. They felt that the set of parameters represented a good compromise between the complexity in their measurement procedures and their ability to discriminate between the three classes reliably across a wide variety of speakers. However, they made it clear that these parameters are not the only possibility and that, quite likely, a better choice could be realised by careful evaluation of a different set.

It is not surprising, therefore, that many different approaches have been used in the past, and will be used in the future, in order to obtain a reliable V-UV classification. Preference of one approach over another is primarily determined by the particular application in which such a system is to be used.

A non-parametric method of classifying speech has

been presented by Seigel[37]. In his method the decision is arrived at by measurement of six features of which no assumption is made about the probability distribution. The design was achieved by viewing the decision-making process as a pattern classification problem and thus defining, (a) a set of speakers from which frames of speech are selected to characterise voiced and unvoiced segments and, (b) a set of features which can be used in making the classification. In order to use as few frames as possible and as few features as possible, a procedure was developed in which the contributions from the two sources of information were interleaved.

Another technique, which uses a different approach from that of Atal and Rabiner and claims simplicity, has been realised by Knorr[38]. In his technique the decision is made by measuring the short-time spectral energy distribution of a low-pass filtered (voiced) and high-pass filtered (unvoiced) speech waveform. The cut-off frequencies of the low-pass and high-pass filters is 1000Hz and 5000Hz respectively. The frequency band between these cut-off frequencies is unused because of ambiguities that exist for certain elements of voiced speech which have sufficient energies above f = 1000Hz, such as higher formant frequencies and high-frequency bursts during pitch excitation. In addition, there are many unvoiced speech segments,

mainly fricatives, which have substantial spectral energies below f = 5000Hz. The simplicity of this technique is evident enough, but its suitability is limited to applications of wideband speech analysis only (ie, > 5000Hz).

The actual effectiveness of these methods has been measured in two ways:

(i) by comparing the number of segments manually classified into the three categories and the number of segments analytically classified by the respective algorithms;

and

(ii) by informal auditory evaluation of the V/UV classifications made by synthesising speech from the analysis parameters and the classifiers V/UV decisions.

3.1.2 The hybrid-TES method

Although all the above techniques are available and have been, individually, shown to be successful, they cannot be readily applied to the task at hand. This chapter therefore presents a technique which is specifically designed for application to TES. Consequently, there are some important differences between this and the previous techniques. These differences are:

(i) The algorithm operates on epochs (or real-zero intervals) rather than on segments of fixed duration. There are two reasons for this. Firstly, in the speech waveform, there are significant acoustic features which last for only one or a few pitch cycles in duration. If the cycles are averaged over a fixed period of time, which is the case in previous techniques, then the information is irretrievably lost. Such transient events frequently occur at vowel-consonant and consonant-vowel boundaries as well as between other acoustically distinct regions, within stop consonants for example[39]. Secondly, this is a feature of time encoded speech, and thus the measurement of it does not incur any extra loading on the processor.

(ii) Strictly, for the hybrid-TES strategy proposed in the previous chapter, it is necessary to differentiate only between voiced speech and unvoiced speech. However, it has been shown[23] that further reductions in the transmission rate requirements can be achieved by encoding the silence regions of speech more efficiently. For this reason the algorithm identifies the three different regions: voiced speech, unvoiced speech and silence.

(iii) Different detection strategies are employed at the different transition regions. Since it is the stable regions of unvoiced sounds which give rise to the high epoch rate the transition regions between the voiced and unvoiced sounds are classified as voiced. That is, the class is not considered to be changed until the speech waveform is well into the unvoiced regions, whereas at the unvoiced to voiced transition the class is immediately changed over.

The transition regions between silence and unvoiced sounds are classified as unvoiced. This is because during the reconstruction process the silence regions are to be reconstructed as silence and if the unvoiced transitions are identified as silence the important unvoiced sound identification cues will be lost.

The criterion used for decision-making is based on two parameter measurements, and the trend of previous such measurements. These parameters are simple to derive and yet highly effective in differentiating between the three classes: voiced sounds, unvoiced sounds, and silence. They are:

(i) The length of time between the real-zero crossings, ie the epoch duration, ED.

(ii) The amplitude of the epoch, EA.

The choice of these parameters is based partly on the application of the algorithm and partly on the knowledge[36] that these parameters vary from one class to another.

3.2 Measurement of Parameters

Figure 3.1 shows the block diagram of the analysis and decision algorithm.

The epoch duration is determined by detecting the occurrence of real-zero crossings and measuring the length of time between successive crossings. The criterion used is to detect a sign change in the sample values and measure the length of time which elapses before the next sign change occurs. First-order linear interpolation enables determination of the zero-crossing position between the two samples of opposite polarity.

The amplitude of the epoch is given by the maximum sample value between the real-zero crossings.

The trend of parameter measurements is determined by keeping a record of the previous epoch amplitudes and durations.

Before proceeding to a detailed discussion of the decision algorithm, it is necessary to discuss observations on the nature of variation of the above parameters for each of the three classes.

Unvoiced speech (fricatives and stop consonants) is produced due to excitation of the vocal tract by a noise-like source at a point of constriction in the interior of the vocal tract. While the spectrum of the noise source is flat, the vocal tract response usually varies with frequency[36]. The fricative sounds are characterised as sustained high frequency regions while the stop consonants usually have a pause portion followed by a high frequency burst.

Time-domain analysis by Baker[39] has revealed that at the beginning and at the end of the high frequency portion of the fricative, there usually is a discontinuity and often a decrease in amplitude where the fricative is preceded by a vowel. At these places is found one or a few cycles characterised by lower cycle-frequencies than those of the other cycles in the acoustic segment immediately preceding and the acoustic segment immediately following this transitional phenomena. Figure 3.4 illustrates this effect in the transition between the sounds 'l' and 's' of the utterance "Charles". As the waveform transitions from prior context into an unvoiced

stop-consonant the cycle-frequency and amplitude drop sharply. This is the familiar pause portion of the stop-consonant. Although this pause portion lasts only one or a few cycles, the cycle-frequencies are quite low, often less than 100Hz, which is consistent with the relatively long duration of this acoustic segment. Figure 3.5 illustrates this in the transition from the sound 'o' to 't' of the utterance "Bottleneck".

Confirmation of some of the remaining observations is provided by the measurements in Chapter 2. The probability distribution shows that the high frequency portions of the unvoiced sounds result in small duration epochs - typically less than 400μs. The epoch rate-v-time graph shows that the high frequency and thus the high epoch rate regions are sustained over quite long periods.

Voiced speech is produced as a result of excitation of the vocal tract by the periodic flow of air at the glottis. The spectrum of the glottal air flow decreases with frequency at approximately 12dB/octave[36], thereby producing a concentration of energy at low frequencies in the speech signal. Consequently, the epoch durations are expected to be large. Indeed, confirmation of this is provided by the probability distribution measurements, where the voiced epochs are, typically, 1.0-2.0ms in duration.

It is difficult to characterise silence[36] (or inter-word spaces) since the parameters can vary considerably from one environment to another, reflecting the variable characteristics of the background or room noise. Quite often, the spectrum of room noise is concentrated at low and middle frequencies. In such cases, the epoch durations for silence would be expected to be larger than for unvoiced speech, but quite comparable to those for voiced speech. In a quiet background the power and hence the amplitude of the silence epochs is lower than for voiced and unvoiced speech.

An interpretation of all these observations and the trend of variation of the parameters for the three classes is given in Figure 3.2. A real scatter diagram for the speech file CBONLY illustrates this trend in Figure 3.6. The decision algorithm makes use of this trend of variation of the parameters and attempts to differentiate between the three classes of speech.

3.3 The Decision Algorithm

The performance of the algorithm can be conveniently expressed by a state diagram, as shown in Figure 3.3. When the algorithm is first initiated the class of speech is silence. Epoch parameters are then measured and compared with the previous measurements (if any) and the class is

revised accordingly. Voiced and unvoiced sounds are differentiated on the basis of epoch durations, whereas the decision between silence and non-silence is made by comparing epoch amplitudes.

Figure 3.7 shows the flow diagram for the decision algorithm. The value of ICLASS determines the class of speech to which the epoch under examination belongs: ICLASS = -1, indicates an epoch in the unvoiced region; ICLASS = 0, indicates silence; and ICLASS = 1, indicates an epoch in the voiced region. When the algorithm is in one of the three states comparisons are made to detect the change of trend and hence revise ICLASS. The trend is determined by monitoring amplitudes and durations of previous epochs over different intervals.

The actual requirements which have to be satisfied at the transition region from one class to another are summarised in Table 3.1 and are described in more detail below.

3.3.1 The silence state

When in the silence state detection of voiced or unvoiced sounds is achieved by comparing amplitude levels, EA, of the epochs with a threshold level, T. Since the decision to change the class has to be made on the trend of epoch amplitudes, it is not sufficient to decide on the

Present state	Next State	Conditions	No of consecutive epochs satisfying the conditions
Silence	Voiced	(EA>T).AND.(ED>DUR)	2
	Unvoiced	(EA>T).AND.(ED<DUR)	2
	Silence	(EA<T)	1
Voiced	Silence	(PAV<T)	LIM
	Unvoiced	(ED<DUR)	LIM
	Voiced	(ED>DUR).AND.(PAV>T)	1
Unvoiced	Voiced	(ED>DUR)	2
	Silence	(PAV<T)	LIM
	Unvoiced	(ED<DUR).AND.(PAV>T)	1

TABLE 3.1 Summary of the state transition requirements

amplitude variation of just one epoch. At the same time however, the decision cannot be delayed for too many epochs because the important transient information may be lost. For this reason, it was decided to take the amplitude information of two successive epochs.

When the amplitude, E_A, of two successive epochs exceeds the threshold, T, the class of speech changes to either voiced or unvoiced, depending on the epoch durations. It changes to voiced if the durations are longer than the value 'DUR' and to unvoiced if they are shorter. Figure 3.8

shows the flow diagram for decision-making in the silence state.

3.3.2 The voiced state

Whilst in this state it is necessary to be able to detect silence or unvoiced sounds. As the voiced sounds are the most informative part of speech, it is important that they remain unprocessed and undistorted. For this reason, a certain amount of hysteresis is incorporated in the algorithm so that the change of class only takes place when the speech is well into the new region. This is accomplished by monitoring the trend of amplitude and duration for a large number of epochs.

When the average amplitude level, PAV, of the previous 'LIM' epochs falls below the threshold, T, the class is changed to silence. The present value, $PAV_{(k+1)}$, is determined from the previous value $PAV_{(K)}$ as:

$$PAV_{k+1} = \frac{(PAV_K * LIM) + EA}{LIM + 1} \qquad \ldots (3.1)$$

where LIM = the number of previous epochs over which the average is taken

EA = the latest epoch amplitude

The initial value of PAV is set to the epoch amplitude before the voiced state is first entered.

Detection of the unvoiced sounds is made by counting the number of consecutive epochs having durations less than the value 'DUR'. When the number of epochs exceeds the value LIM the class is changed to unvoiced. The counting procedure is a necessary step in differentiating between small amplitude, small duration epochs in the voiced region and the true unvoiced epochs. Figure 3.9 shows the flow diagram for this state.

3.3.3 The unvoiced state

Here, the requirement is the detection of silence or voiced sounds. Silence is detected by monitoring the average epoch amplitude level, as in the previous case. Detection of voiced sounds is made by monitoring the duration of epochs. When the duration of two consecutive epochs exceeds the value 'DUR' the state is changed to voiced. The flow diagram for this state is given in Figure 3.10.

3.4 Determination of the Reference Parameters

Successful operation of the algorithm depends on the values used for the reference parameters, which are:

(i) the epoch duration boundary, DUR, between the voiced sounds and the unvoiced sounds;

(ii) the threshold amplitude, T, for differentiating between silence and non-silence;

(iii) the number of consecutive epochs, LIM, for differentiating between true unvoiced epochs and the small amplitude, small duration epochs within the voiced regions.

Due to the lack of theoretical concepts for analytic evaluation, the parameters are determined by experiment.

The decision algorithm is incorporated into the program, CLASS, and is used to process speech files of the utterances listed in Appendix B. The program accepts as input the speech file to be processed and pre-set values for the reference parameters. It outputs a classified file, which has three amplitude levels: positive for voiced speech; negative for unvoiced speech and zero for silence.

The approach for determining the pre-set values of the reference parameters is to process one speech file and determine the most relevant values and to later amend them (if necessary) for the remaining files. Each of the parameters are determined by first presetting two and noting the effect of variation of the third. When the most appropriate value has been selected for this variable parameter, it is pre-set to this value and one of the remaining two are varied. In this way the most appropriate values for all three parameters are determined.

3.4.1 Threshold amplitude, T

The purpose of this reference parameter is to differentiate between the silence and non-silence regions of speech. The one significant difference between these regions is the amplitude of their epochs - being small for silence and large for non-silence. Successful separation can thus be achieved by comparing the amplitude of the speech epochs with a fixed reference value. If the epoch amplitude is less than the reference then it may be regarded as a silence epoch and if it is more than the reference it may be regarded as a non-silence epoch.

The idea of a reference voltage has been introduced, previously, for the measurement of the probability distribution of real-zero intervals of speech, where the requirement was to separate the portion of the probability distribution due to the speech wave alone and that due to the system noise in the inter-word spaces. It was concluded then that the reference (or threshold) voltage should be a variable, dependent on the speech signal to noise ratio. However, the requirement in this case is different. The threshold voltage is used to identify regions of the speech wave which may be reconstructed as silence. Reconstruction of noisy inter-word spaces as silence would result in a very unnatural effect. Thus, for conversations conducted in a noisy background silence does not exist. In such cases the algorithm should only distinguish between the

voiced and unvoiced regions, with the category of the inter-word spaces depending on the epoch durations. This is achieved by having a fixed amplitude reference, which has to be determined with care: too high a value would cause the non-silence regions to be regarded as silence, hence causing distortion during reconstruction, and too low a value would allow much of the silence to go undetected. Since the reason for detecting silence is to allow a reduction in data, it is important that the maximum amount of silence is detected. A measure of the amount of silence detected can be given by the epoch reduction achieved when the individual silence segments are each regarded as one epoch. Thus, determination of the parameter can be made by measuring the epoch count reduction and the distortion introduced with the parameter set to different values, where distortion is measured qualitatively by audible comparison.

The speech file CBONLY is used for initialising the parameters as it comprises a fair proportion of all three classes of speech (voiced, unvoiced and silence) and they are easily identifiable. Figure 3.11 shows the reduction in epoch count against the threshold value for this file, under varying conditions. The graph shows a sharp rise in the reduction of epochs up to a threshold of about 10.0*. Beyond this value the rise in epoch reduction is less steep. Introduction of distortion is apparent at higher

* Threshold amplitude in machine units

values of threshold, although a reduction in the value of LIM causes distortion to be noticed at lower threshold values. With LIM = 40.0 distortion is observed at a threshold of about 40.0, whereas LIM = 20.0 causes distortion at a threshold of about 30.0. Additionally, a reduction in LIM gives higher epoch reductions.

In order to steer clear of distortion and yet still achieve a reasonable reduction in epochs, a threshold amplitude of 20.0 (or 38dB below the peak) was selected. Processing of other speech files with this value of threshold gives a minimum epoch reduction of 10% (APPLE8) and in one case as much as 20% (BIRD).

These figures are very much lower than the figures presented by Al-Doubooni[40], where he showed that epoch rates of three different utterances could be reduced from 2075, 2181, 2163 to 1103, 1004, 704 respectively by symmetrically thresholding at between 30 and 36dB below the peak. These high epoch rate reductions are possibly due to the unusually large epoch rates before thresholding. In addition, the thresholding method used by Al-Doubooni eliminates all epochs below the threshold level and not just in the inter-word spaces.

Indeed, advantage could be taken of these extra large reductions by performing the thresholding, as presented by

Al-Doubooni, before attempting to classify the waveform into the three regions. In this way the silence regions could be more easily identified, since in these regions the amplitude would be zero and therefore below the threshold of the classification algorithm.

3.4.2 Epoch duration, DUR

This parameter is used to differentiate between the epochs which are definitely due to voiced sounds and those which may be due to voiced or due to unvoiced sounds. For instance, if the epoch duration is longer than the reference value, DUR, then the epoch is almost certainly due to voiced sounds. If, on the other hand, the duration is less than DUR the epoch may be due to voiced or unvoiced sounds and further comparisons are necessary for classification.

The value of the reference duration depends on the effect of variation. Too high a value will cause some of the voiced epochs to be regarded as unvoiced epochs, hence causing distortion during reconstruction; and too low a value will make the algorithm very selective, such that only very high frequency unvoiced sounds are detected.

In order to ensure optimum data reduction the maximum length of unvoiced region has to be detected. Again, the epoch reduction achieved, by regarding each

detected segment of unvoiced sound as one epoch, gives an indication of the data reduction. The determination of DUR can thus be achieved by noting the effect it has upon the speech quality and the epoch count.

Figure 3.12 shows the effect on epoch count of varying DUR and LIM; and Figure 3.13 shows the introduction of distortion with variation of DUR and LIM, where the distortion was judged subjectively by three trained listeners.

The speech file CBONLY was processed with different values of DUR and LIM, in which the detected unvoiced regions were replaced with zero amplitude samples. The listeners were presented with these processed files in the order shown in Figure 3.3 and were asked to determine the highest value of LIM at which they could not detect any distortion of the voiced regions of the utterance. One listener judged distortion to be absent above LIM = 15, whereas the other two concluded that distortion was present for up to and including LIM = 20. These values were confirmed when the files were presented in reverse order.

From the point of view of epoch reduction, the higher value of DUR is desirable. However, the higher value of DUR leads to a greater distortion, unless the

value of LIM is maintained high. Since it is desirable to achieve the highest epoch reduction a value for DUR of 4.0 (400μs) was selected. This imposes the restriction that the value of LIM be in excess of 25 for distortion-free classification.

3.4.3 Epoch count, LIM

This parameter is used to differentiate between the small duration epochs which may be in the midst of voiced sounds and those due to unvoiced sounds. The sequence of epochs having duration less than the value DUR are counted. If the number of such epochs is less than the reference value LIM then these epochs are regarded as due to voiced sounds. When the number exceeds the reference value LIM, the epochs are regarded as being due to unvoiced sounds. As explained in the previous section, two effects of varying the value of LIM are: distortion and epoch reduction. A third effect comes to light after a visual observation of the classification of the speech file and this is sensitivity to change of class, where sensitivity is defined as a momentary change of class. Figures 3.12 and 3.13 show the epoch count and distortion effects respectively. Figure 3.14 illustrates the sensitivity effect and lists the blocks of the speech file in which these momentary changes are detected.

Selection of DUR required that LIM should be at least

25. Sensitivity is lowest when LIM is high. Also, high values of LIM give lower epoch reductions. Bearing these restrictions in mind, LIM was set at a value of 30.

3.5 Results Analysis and Discussion

The reference parameters were set to their determined values and the algorithm (CLASS) was used to identify the different regions of speech for the speech files listed in Appendix B. It is both pointless and space consuming to illustrate the results for all speech files. Instead, only the classification for the speech file CBONLY is shown. Figure 3.15 shows the speech waveform with the class identification superimposed. The three classes of speech are indicated by three amplitude levels: positive for voiced; negative for unvoiced; and zero for silence.

In the class identification of Figure 3.15 three areas of sensitivity can be observed, one in block 2 and two in block 69. They indicate difficulty in differentiating between silence and unvoiced sounds. This does not represent a serious problem since if the silence is incorrectly identified as unvoiced then the amplitude information will be such as to make the unvoiced sounds inaudible after reconstruction. If, on the other hand, the unvoiced sounds are identified as silence, then the amplitude of the unvoiced sounds must be so low as to be

considered inaudible. A problem could be encountered if any of the voiced sounds are incorrectly identified as silence or unvoiced. However, this is unlikely because of the low value used for the threshold, T, and the high value for the reference parameter, LIM.

The advantage of a low threshold is that epochs having amplitudes less than the threshold will be inaudible and therefore can be regarded as silence. Epochs having amplitudes larger than the threshold can be regarded as non-silence. If the background noise is such that there are no epochs below the threshold amplitude, the algorithm fails to identify any silence regions, instead it just differentiates between voiced and unvoiced sounds. This is a desirable effect since if the background noise during inter-word or inter-sentence periods is audible it would be unnatural to transmit it as silence. Thus, with a low threshold value a high background noise is identified as either voiced or unvoiced sounds, depending upon which criteria is met.

Processing of the remaining speech files with the reference parameters set at T = 38dB below peak, DUR = 400μs and LIM = 30 shows that the three classes of speech are successfully identified. Here success is defined as 'incurring no errors in the identification of voiced regions', although permitting some sensitivity in the silence or unvoiced regions.

CHAPTER 4

DETERMINATION OF THE ENCODING AND RECONSTRUCTION PARAMETERS FOR UNVOICED SOUNDS

4.1 Introduction

Having designed an algorithm for identifying the unvoiced regions of speech, it is now possible to investigate the encoding and reconstruction aspects of hybrid-TES. It is intended here to replace the identified unvoiced regions of speech by spectrally shaped noise and hence determine the parameters which allow optimum reconstruction.

Initial tests have shown that a single noise source is not sufficient to replace all unvoiced sounds; instead the availability of many noise sources, with different frequency characteristics, is required for good quality reconstruction. Embodied within this requirement is the added task of identifying which of the noise sources must be used to reconstruct a particular region of the unvoiced sounds.

4.1.1 Identification of the Sounds

In essence, the identification problem lies in differentiating between the different unvoiced sounds. Standard English has the following unvoiced sounds: the fricative consonants, /f/ as in "leaf", /s/ as in "base",

/sh/ as in "leash", /th/ as in "teeth", /ch/ as in "church"; and the stop consonants, /p/ as in "pet", /k/ as in "kit" and /t/ as in "ten". Generally, the fricatives are characterised as sustained high frequency regions, whereas the stops are characterised by "silence" preceding a burst of high frequency signal. When the stop is adjacent to a vowel, there are usually three regions: transition, silence and burst or silence, burst and transition; where the transition is the rapid change in the formant frequencies. Hughes, Halle and Radly[41] have shown that of these three regions only the silence is a necessary cue - the silence with either transition or burst is a sufficient cue for identifying a stop. However, except for the differentiation between /f/ and /th/ the transitions of the formants in the adjacent vowels contribute little towards the identification of the fricatives. Thus, the primary cues for identification of the fricative sounds are contained in the spectra of the fricative portion.

Hughes et al[42] have determined these spectral properties of the unvoiced sounds and consequently have developed spectral analysis techniques for identifying these sounds individually. However, since one requirement of TES is low terminal cost real-time identification of the sounds through spectral analysis is not feasible. On the other hand, Ito and Donaldson[43] have shown that unvoiced sounds are distinguishable from each other on the basis of the

zero-crossing (or epoch) rates, the determination of which is a simple task. They manually selected 10ms segments from stable regions of the unvoiced sounds and performed measurements of zero-crossings for the normal signal and for the differentiated signal. They concluded that the unvoiced fricatives /s/, /sh/ and /f/ are differentiable from each other, solely on the basis of the zero-crossing rate of the derivative signal, whereas only the sounds /s/ and /sh/ can be distinguished by the normal signal. In addition, for unvoiced stop consonants the zero-crossing rate of either the normal signal or its derivative is useful for classification, provided some information concerning the contextual environment is available.

These measurements were performed on broad-band speech which is bandlimited between 0-8kHz. Since some elements of voiced speech (eg higher formant frequencies and high frequency bursts during pitch excitation[32]) have sufficient energies up to about 4000Hz, differentiation succeeds in further separating the voiced and unvoiced sounds. However, for severely bandlimited speech (eg 300-3400Hz) differentiation of the signal only helps to equalise the zero-crossing rates of the higher frequency voiced sounds and the unvoiced sounds, thus making it difficult to distinguish between them.

For the present case, therefore, identification of the

unvoiced sounds is attempted on the basis of the zero-crossings of the normal speech. Since the zero-crossings alone are not sufficient to identify individual sounds, additional cues are provided by the classification algorithm which correctly codes the silence and transition regions of the stops; and the first 30 epochs (≅ 30 x 0.4 ≅ 12ms) of the fricative sounds.

4.1.2 Encoding and reconstruction

Another important requirement of TES is that the catalogue of symbols* must be kept at a minimum. Hybrid-TES obviously increases the number of symbols in the catalogue because it requires additional symbols to describe the unvoiced sounds. However, the increase can be kept to a minimum if the number of symbols used to describe each noise source is restricted to one, ie if the number of noise sources required is eight, then only eight extra symbols are needed in the catalogue for hybrid-TES.

This requirement can be satisfied by encoding the complete unvoiced region in short segments of fixed duration, where the average epoch rate during each segment determines the noise source to be used during reconstruction. In addition, this segmentation procedure allows closer tracking of the spectral characteristics of the unvoiced sounds.

* Refers to encoded epochs

For unvoiced regions which are shorter than the segment duration reconstruction can be achieved by using voiced symbols. This is an adaption from the vocoder technology.

In a vocoder it is required to differentiate between voiced and unvoiced sounds. The decision is based on speech segments of 10ms duration. If the 10ms segment contains a transition between unvoiced and voiced sounds and the unvoiced portion is greater than half, then, depending on the parameters used for decision-making, the segment could be labelled as voiced. This means that the unvoiced portion will be reconstructed as voiced. The speech reconstructed in this way is of acceptable quality. Therefore, it seems feasible, for hybrid-TES, to replace portions of the unvoiced sound which are less than the segment duration, by voiced symbols of the same duration, without incurring excessive distortion.

In the previous chapter it was pointed out that silences can be transmitted more efficiently than by simply encoding between the real-zero intervals. In this chapter an actual method for encoding silences is proposed, which is to encode it similarly to the unvoiced sounds, ie by segmentation of the silence regions. The only difference between encoding unvoiced sounds and silences is in the amplitude information, where for silence segments the

amplitude information is zero. In this way encoding of silences does not incur any additional symbols in the catalogue.

4.2 Segmentation of the Identified Regions

Since encoding of the unvoiced and silence regions of speech is to be achieved by segmentation, an algorithm is designed for this purpose.

The identification algorithm determines the class of each epoch individually. Depending on the class, the epochs are processed in one of three ways.

4.2.1 Unvoiced epochs

The duration of each of the unvoiced epochs is added to obtain unvoiced segments of fixed duration. A record is also kept of the number of epochs required to make up the segment. This allows calculation of the average epoch rate, which is the parameter that is used during the reconstruction of the segment.

For regions where the duration of the unvoiced sound is not an exact multiple of the segment duration, the end portion is encoded as a voiced symbol. When the epoch class changes to silence or voiced the stored unvoiced duration is processed first. Figure 4.1 shows the flow diagram.

4.2.2 Silence epochs

Segmentation of the silence regions is made in exactly the same way as the unvoiced regions, with the exception that the amplitude information accompanying the encoded symbol is zero. In addition, the remaining information (if any) from the previous class of epochs is processed before the silence epochs. Figure 4.2 shows the flow diagram for segmenting the silence regions.

4.2.3 Voiced epochs

Since the voiced regions are encoded epoch by epoch there is no need for segmentation, and as such no processing is required. The only point to note here is, again, the requirement for encoding the remaining stored information from the previous class. Figure 4.3 shows the flow diagram.

4.3 Substitution of the Identified Regions

Adequate substitution of the unvoiced segments by noise forms the crux of the feasibility study. The underlying factors affecting the quality of reconstruction are:

(i) the identification of the different unvoiced sounds;

and (ii) the availability of numerous noise sources for reconstructing these different sounds.

As regards the identification, the only means available, within the constraints of TES, is through detection of the epoch rate variation due to the different sounds.

The limitation with regard to the availability of noise sources is the size of the symbol catalogue, which is to be kept at a minimum.

During substitution consideration should be given to the amplitude and spectral information of the segment. A simple means of matching the amplitudes is through evaluation of the rectified mean of the unvoiced segment and modifying the noise source amplitude accordingly. Spectral matching offers two possibilities - average epoch rate, and spectral shaping.

4.3.1 Average epoch rate matching

The basis of this method is derived from the work by Heinz and Stevens[44] on the generation of unvoiced sounds from bandlimited noise. They analysed the spectral aspects of the fricative consonants from the results presented by Hughes et al[42] and attempted to generate similar spectral responses from noise-excited electric circuits whose pole and zero frequencies could be varied. They showed that reasonably good approximations to the fricative spectra of the sounds 's', 'sh' and 'f' could be obtained by white-noise excitation of circuits characterised by one zero (conjugate-pair) and one pole. After perceptual testing the generated sounds were characterised as resonances in the range: 2200Hz - 2700Hz for 'sh'; 3500Hz - 6400Hz for 's'; and 6500Hz - 8400Hz for 't' and 'th', where the

transitions distinguish between the two sounds.

To the author's knowledge no similar study has been conducted for the stop consonants, although it has been shown in the past[28-31] that the burst region of the stop is similar in characteristics to the fricative sounds. Since it is the burst region of the stop which is to be reconstructed in hybrid-TES, it is not too unreasonable to assume that it could be regenerated similarly to the fricative sounds, ie by bandpass filtering random noise.

In a paper on the mathematical analysis of random noise Rice[45] has shown that filtering of noise effectively controls its zero-crossing rate. He has related the expected number of zero-crossings per second, Z, to the upper, fb, and lower, fa, cut-off frequencies of an ideal band-pass filter by the equation,

$$Z = 2\left[\frac{1}{3} \cdot \frac{fb^3 - fa^3}{fb - fa}\right] \qquad \ldots (4.1)$$

This method of substitution thus relies on matching the average epoch rate of the unvoiced segment to the expected zero-crossing rate of the band-pass filtered noise source. Since there are eight unvoiced sounds a simple interpretation of this method requires that eight band-pass filtered noise sources be available to emulate the different sounds.

4.3.2 Spectral shape matching

From equation (4.1) the expected number of zeros per second for a noise source bandlimited between 2.0kHz and 3.0kHz is 5000. In addition, the expected number of zeros per second for a noise source bandlimited between 1.0kHz and 3.75kHz is 5000. However, the sounds produced after these two filtering operations are quite different. The same applies to the unvoiced sounds in that the average epoch rates of two sounds may be similar, but because of their different spectral shaping the sounds are quite different. This illustrates that the sound associated with a given zero-crossing rate depends on the centre frequency of the filter as well as its bandwidth. Consequently, an alternative method of substitution may be considered. In this method the noise source is spectrally shaped, in accordance with the unvoiced sound spectrogram, instead of simply being band-pass filtered, to match the epoch rate.

Actual spectrograms of the speech files CBONLY, BIRD, APPLE8 AND APPLE7 are available and they contain all the unvoiced sounds except the sound 'f'. However, Potter et al[46] in their book "Visible Speech" provide a comprehensive list of the typical spectrograms of all the different sounds. A summary of these, for the unvoiced sounds, is given in Figure 4.4, and the author's estimation of the filter characteristics required to generate them from noise is shown in Figure 4.5

4.4 Determination of the Parameters

The segmentation and substitution algorithms are both implemented in a program, RECON9, to determine the quality and parameters of reconstruction. The program accepts as input the original speech file, the classified file and up to nine noise files; and outputs one reconstructed speech file.

The classified file identifies the three different regions of the original speech file. Each region is represented by a different amplitude level - positive for voiced; negative for unvoiced; and zero for silence.

The original speech file is used to provide the necessary information for reconstruction. For unvoiced regions it provides the amplitude and average epoch rate information and for voiced regions it is available for direct copying into the reconstructed file.

The different noise files are used for reconstructing the different unvoiced sounds. Each noise file is generated by filtering the noise source with the required spectral characteristics. The use of noise files is preferred to generating noise as required because it simplifies programming and enables faster processing speeds. In addition, it makes the program flexible to any changes in the method of substitution. Figure 4.6 shows the flow diagram for the program.

After initialisation the first blocks of all files are read into the appropriate arrays. Inspection of the individual sample amplitudes in the classified array determines the action to be taken. If the amplitude is negative 'I' samples of the original speech file are replaced by 'I' samples from one of the noise files, where 'I' corresponds to the encoding segment duration. The choice of noise file is determined by the number of zero-crossings detected in the 'I' samples of the original speech file. Should the classified amplitude change before an interval of 'I' samples, then the original samples are replaced by zero. These shorter than segment duration regions of the unvoiced sounds should ideally be replaced by voiced symbols of the appropriate duration. However, since the object of this investigation is to replace unvoiced sounds by noise without incurring excessive distortion, TES encoding at this stage only makes it difficult to isolate the proportion of distortion due to the substitution process. Thus, to avoid the effect of TES quantisation the shorter than segment duration regions are replaced by zero amplitude samples.

If the sample amplitude in the classified array is zero the original samples are replaced by zero, and if it is positive the original samples remain unaltered.

Execution of the program allows evaluation of the

quality of reconstructed unvoiced sounds and in turn the determination of parameters for the highest quality, where the parameters to be determined are: the characteristics of the reconstruction noise and the duration of the encoding segment. The determination of the highest quality is achieved by repetitive processing and comparison of reconstruction with different parameters. Because the number of repetitions may be large, a single trained listener is employed for continuous assessment of the quality. However, confirmation of the results is provided through gross comparisons by many trained listeners, as explained below.

4.4.1 Characteristics of the reconstruction noise

The noise characteristics are determined by generating the noise files in accordance with the spectral requirements of the substitution method and executing program RECON9 to achieve optimum quality.

A straightforward interpretation of the average epoch rate method of substitution requires that the eight noise files cover the complete frequency range of the unvoiced sounds, which in this case can extend from about 500Hz to 4.5kHz. Figure 4.7 shows the subdivision of the unvoiced frequency range and its corresponding epoch rates, between the eight noise files. The reconstructed speech is fully intelligible, though not as good in quality as the original.

When attempting to confirm that the reduction of the number of noise sources gives poorer quality reconstruction, it was observed that the degradation of quality could not be perceived until the number of noise files was reduced to below four. The approach adopted was to reduce the number of files one by one (as illustrated in Table 4.1) and reconstruct the speech files CBONLY and BIRD; and compare their quality with the initial attempt at reconstruction. Out of the four trained listeners used for gross comparisons, only one was able to perceive a difference between the eight source reconstruction and the four source reconstruction, whereas all four listeners agreed on the difference between the eight source and the three source reconstruction. Consequently, it was decided to use four noise files for subsequent reconstruction, the parameters of which are shown in Figure 4.8.

Initial Attempt	1	2	3	4	5	6	7	8
2nd Attempt	1		3	4	5	6	7	8
3rd Attempt	1		3		5	6	7	8
4th Attempt	1		3		5		7	8
5th Attempt	1		3		5		7	
6th Attempt		1			5		7	
7th Attempt		1				5		

——————————→
number of noise files

TABLE 4.1

The quality of the reconstructed sounds can possibly be improved if, instead of being simply bandpass filtered, the four noise files are spectrally shaped, in accordance with the spectrograms of the sounds they are replacing. Fortuitously, the seven filter characteristics in Figure 4.5 can be separated into the three categories: low-pass, band-pass and high-pass. In this way the filter order and cut-off frequencies for each category can be designed for the sharpest sounding reconstruction of the group of sounds which it replaces. The sounds 'k' and 'p' can be regarded as low-pass; the sound 'sh' as band-pass; and the sounds 's', 't', 'th' and 'f' as high-pass; whereas the sound 'ch' can be generated from a combination of the sounds 't' and 'sh'. Thus, the four noise files can be spectrally shaped as:

noise file (i) - low-pass
noise file (ii) - band-pass
noise file (iii) - high-pass
noise file (iv) - high-pass

Measurement of the epoch rates for the different unvoiced sounds can be made by program RECON3. This program requires as input: the original speech file, containing the unvoiced sound, and the classified version of the original speech file; and it outputs a results file which presents the number of zero-crossings detected in successive segments of the unvoiced and silence regions of the original speech file. The results of measurement on the unvoiced

sounds detected in the utterances of Appendix B are shown in Figure 4.9. These results, however, are not general enough to base on them a speaker independent reconstruction strategy. More characteristic results have been presented by Ito and Donaldson[43] (Figure 4.10), although they have been measured on broad-band speech. It is shown in Chapter 5 that the effect of filtering is to alter the epoch rates. Bearing in mind that the utterances to be reconstructed, in this case, are bandlimited between10-4.5kHz, the results presented by Ito and Donaldson can be applied here by rescaling the axes. Such a process enables the different unvoiced sounds to be conveniently separated into four epoch rate groups, as:

Group (i)	< 2800	for sounds k, p	
Group (ii)	2801 < 4800	for sounds sh	ch
Group (iii)	4801 < 6800	for sounds f, th, t	ch
Group (iv)	6801 < 8800	for sounds s	

Thus, the four epoch rate groups can be matched to the four noise files, the characteristics of which are determined by experiment, for the sharpest sounding reconstruction of the unvoiced sounds.

Figure 4.11 shows the spectral characteristics for the above requirements, where the experimental approach adopted was to arbitrarily set the filter characteristics for the four files at some initial value and to adjust each one, in turn, to achieve the sharpest reconstruction of the sounds

it replaces. The group (i), group (iii) and group (iv) sounds were adequately reconstructed by the types of filter characteristics predicted in Figure 4.5. However, optimum quality for the group (ii) sounds could only be achieved by means of a second order low-pass Chebyshev filter with cut-off at 3.3kHz. This type of characteristic is approximately equivalent to the band-pass characteristic predicted in Figure 4.5, but with reduced attenuation in the stop band. Figure 4.12 shows the gain response of the four filter characteristics for the sharpest reconstruction. Gross comparison by four trained listeners confirmed that the reconstruction using the characteristics of Figure 4.11 is sharper than the reconstruction using the characteristics of Figure 4.8.

4.4.2 Segment duration

Encoding of the unvoiced regions by segments of fixed duration gives rise to a problem when the unvoiced sound duration is not an exact multiple of the segment duration. The end portion, which can vary from zero to very nearly the segment duration, requires the availability of many unvoiced symbols, to encode all possible durations. On the other hand, TES requires that the symbol catalogue be kept to a minimum. A solution to the problem could be sought through encoding the end portion by a voiced symbol. This, inevitably, introduces distortion into the speech waveform, the amount of which depends on the value of the fixed segment duration.

Guidelines for fixing the value are drawn from the wealth of literature on vocoders, where it is necessary to determine the type of source to be used at the synthesis end. The decision is made, at the analysis end, by measuring parameters of successive 10ms segments. Depending on the parameters, the 10ms segment, containing half unvoiced and half voiced speech, is often labelled as voiced. In such cases, the unvoiced portions (≅5ms) are reconstructed by a buzz source, obviously incurring distortion. However, the quality of speech is readily accepted. In the present case, therefore, it seems feasible to set the segment duration to 5ms.

Actual experiments were conducted with the durations set at 5ms and 10ms. The increased distortion was easily noticeable with the larger value. It was thus decided to fix the duration at 5ms.

4.5 Conclusions

Although the unvoiced sounds can be generated easily enough from random noise by spectrally shaping the filter characteristics, identification of the individual sounds is not possible by epoch rate alone, which only succeeds in identifying one of four groups of unvoiced sounds. There appears to be no advantage in using more than four noise sources for reconstruction.

Processing of the available speech files, with only four noise sources, retains the intelligibility of all the unvoiced sounds although some naturalness is lost as compared with the original. This confirms the view that the noise sources only approximate the spectral identification cues, whereas the remaining important identification cues are provided by the classification algorithm.

The number of extra symbols required in the catalogue for hybrid-TES can be restricted to four by encoding the unvoiced and silence regions in segments of 5ms duration.

CHAPTER 5

PERFORMANCE ANALYSIS OF HYBRID-TES

5.1 Introduction

The feasibility study is concluded by measuring the performance of the hybrid-TES encoding technique, in comparison with ordinary TES behaviour.

Hybrid-TES is designed to enable savings in the transmitter and receiver buffer size, and delay. A consequence of this encoding method is the requirement of the addition to the catalogue of four extra symbols.

It is intended in this chapter to quantify the various savings and to determine the overall advantage or disadvantage of the technique, bearing in mind the overheads that are necessarily incurred. Comparisons are made on the basis of the buffering requirements, the probability distributions and the quality of reconstructed speech.

An investigation is also made of the effect of signal processing on hybrid-TES.

5.2 Buffering Requirements

TES is an example of a variable date-rate source encoding scheme. Matching of the variable source-rate to the constant transmission rate of a communication system

is achieved by inserting buffer stores at the transmitter and at the receiver. An inherent consequence of this buffering strategy is the introduction of a delay into the communication system. In a recent paper by Turner et al[47] it is shown that fairly large delays (≅ 1 sec) may be expected in the distortionless transmission of TES. These delays are such as to make two-way conversation difficult and it has been concluded that further redundancy removal techniques have to be developed to remove the large variations that can occur in basic zero-crossing rates.

Hybrid-TES may be regarded as one such technique. It enables a reduction in the requirement of the delay and buffer size when compared with ordinary TES.

5.2.1 Delay

The requirement of a system delay arises as the result of buffer underflow. That is, the situation in which a request is made to the buffer for information, when in fact the buffer is empty. Because of the variable nature of the data source and the constant transmission rate, buffer underflow may occur frequently, both at the transmitter and at the receiver. At the transmitter it occurs during the relatively long-term persistence of data rates lower than transmission rate. At the receiver it occurs during periods when the symbol rate is higher than transmission

rate. Both these conditions of underflow can be avoided by the introduction of a delay before information is taken from the buffer. Figure 5.1 illustrates the concept.

The system delay is the sum total of the delays required at the transmitter and at the receiver. An alternative to providing a fixed delay at the transmitter, to avoid buffer underflow, has been suggested in (47) as the 'symbol injection' procedure.

Symbol injection begins when either the transmitter buffer is empty, or when the rate at which information is required from the buffer is greater than the rate at which it enters the buffer from the source. At the receiver these injected symbols are simply discarded. The system delay, in this case, is a function of the delay required at the receiver and the delay introduced due to symbol injection. Expressions are derived in (47) which estimate the minimum system delay required.

From (47), the minimum system delay, β, for distortionless transmission is given as:

$$\beta \geq \frac{[N - R + S_j]_{max}}{T_r} \quad \text{for all } t > 0 \qquad \ldots (5.1)$$

where N is the cumulative total number of symbols entering the buffer up to time t

R is the instantaneous rate of source symbol generation

S_i is the number of symbols injected up to time t

T_r is the constant transmission rate

β can be divided into its delay at the receiver, Δ, and delay due to symbol injection, δ, so that:

$$\beta = \Delta + \delta \qquad \ldots (5.2)$$

where $$\Delta = \frac{[N - R]\ \max}{T_r} \qquad \ldots (5.3)$$

and $$\delta = \frac{[S_i]\max}{R} \qquad \ldots (5.4)$$

The delay at the receiver, Δ, is a function of the source, the transmission rate and the number of symbols in the buffer. Since the source symbol generation cannot be controlled, a reduction in the receiver delay can only be achieved in one of two ways: (i) by increasing the transmission rate, T_r, as suggested by Mason and Balston[48]; and (ii) by reducing the number of symbols, N, in the buffer. It should be noted, however, that reduction of N beyond R does not necessarily cause further reduction in the system delay. This is because symbol injection begins when $N < R$ and a corresponding increase is caused in δ, the delay due to symbol injection.

Hybrid-TES encoding reduces N by identifying high data rate (unvoiced) regions of the speech waveform and replacing them by fewer symbols of much longer duration. The actual reductions depend on the relative proportions

of the duration and emphasis of unvoiced sounds. That is, utterances with long, highly emphasised unvoiced sounds achieve the highest reductions in the delay requirements.

5.2.2 Buffer size

The basic requirement of the buffer is that it must be capable of storing the parameters of each of the symbols used to encode the speech waveform. The size of the buffer, defined as the capacity of symbol stores, is a function of the symbol rate, N, and the transmission rate, T_r. Since the transmission rate remains constant the buffer size depends solely on the symbol rate. The variable nature of the symbol rate imposes variable requirements on the buffer size.

Various repeat strategies are being considered elsewhere[49,50] which permit a maximum buffer size to be defined. Basically, when the symbol rate exceeds the requirements of the buffer size limits, the input to the buffer is discarded and a repeat symbol is stored. When transmitted, the repeat symbol instructs the receiver to repeat the previous 'X' symbols when decoding. 'X' corresponds to the duration for which the input to the transmit buffer was discarded. Obviously, the reconstructed waveform becomes distorted and the amount of distortion depends on the number of repeats.

Reduction of the symbol rate obtained by hybrid-TES encoding, leads to a reduction in the buffer size requirements. Consequently the buffer size limit is invoked less frequently. Thus, the number of repeats and hence the amount of distortion in the reconstructed waveform is reduced. Actual reductions depend on the proportions of the utterances for which the symbol rate can be reduced.

5.2.3 Comparison of the delay and buffer size for TES and hybrid-TES

The delay and buffer size requirements for TES and hybrid-TES encoding can be determined by using the available buffer simulation program, BCRSIM, the basic form of which is illustrated in the block diagram of Figure 5.2.

The digitised speech file is presented to the input of the system, which measures the real-zero intervals, and stores the information in the transmit buffer. The information is subsequently transferred to the receive buffer, at regular intervals (the transmission rate). After a delay, β, the system begins to reconstruct the speech waveform. The zero-content detector monitors the state of the transmit and receive buffers. When the transmit buffer is empty or low, symbol injection is initiated. When the receive buffer empties the program is stopped, implying insufficient system delay, β. When the program terminates, for whichever reason, the output terminal prints all the system parameters.

Evaluation of the minimum delay required in transmitting a speech file without distortion is made by trial and error. A low value of delay is set and the program executed. If the program is completed the delay is adequate and a further reduction in delay may be attempted. If the program stops when incomplete, then the delay has to be increased and the program re-executed. This process is repeated until the minimum value for the delay is obtained.

The input to the system has to be of the form which allows determination of the real-zero intervals. Thus, when performing the measurements for hybrid-TES encoding care has to be taken to ensure that the input is compatible with the system requirements. This is achieved by program, TESRE1, which replaces the unvoiced and silence regions of the speech file by alternate polarity parabolic type symbols, of fixed amplitude and duration, as shown in Figure 5.3.

Two series of measurements are made: firstly, the variation of delay and buffer size with transmission rate, and secondly the variation of the number of repeats with different buffer size. The transmission rate and the minimum delay, for the latter measurements, are preset to the values determined by the first series of results. Utterances listed in Appendix B are used as the test

material and the results of measurement are presented graphically.

Figures 5.4-5.8 show the variation of the delay with the transmission rate, of TES and hybrid-TES superimposed, for the five different utterances; and Figures 5.9-5.13 show similar results for the variation of buffer size with transmission rate. Figures 5.14-5.18 show the number of repeats, and hence an indication of the distortion introduced into the reconstructed waveform, with varying buffer size. For the latter case, the transmission rate was prefixed at 1000 symbols/sec, except for utterance CBONLY which was set at 1250 symbols/sec, and the delay set to the minimum value required for ordinary TES to just avoid receiver buffer underflow for each of the utterances.

5.2.4 Result analysis and discussion

Examination of the graphs in Figures 5.4-5.13 shows that considerable reductions are possible in the delay and buffer size requirements by using hybrid-TES encoding. Table 5.1 compares the possible reductions at three different transmission rates. It is observed that the percentage reductions are approximately similar for the three transmission rates. Alternatively, the reductions may be considered in terms of transmission rate, that is, fixing the delay at a preset value and noting the reductions possible in the transmission rate. Table 5.2 shows

	DELAY			BUFFER SIZE			
Utterance	% reductions in delay due to Hybrid-TES			Utterance	% reductions in buffer size due to Hybrid-TES		
	At 800 symbols/ sec	At 1000 symbols/ sec	At 1250 symbols/ sec		At 800 symbols/ sec	At 1000 symbols/ sec	At 1250 symbols/ sec
FEM1	17	23	32	FEM1	17	24	33
APPLE8	78	74	63	APPLE8	75	74	81
APPLE7	86	86	81	APPLE7	85	86	80
BIRD	90	97	97	BIRD	86	94	96
CBONLY	-	-	86	CBONLY	-	-	86

TABLE 5.1

Utterance	% reduction in transmission rate due to hybrid-TES
	Delay = 200ms
FEM1	8%
APPLE8	21%
APPLE7	35%
BIRD	51%
CBONLY	70%

TABLE 5.2

these reductions with the system delay fixed at 200ms, which is considered to be tolerable[48].

Excepting the results for utterance FEM1 for the time being, as requiring special attention (Section 5.4), the lowest reductions achieved are for utterance APPLE8. It is observed from Table 5.2 that the transmission rate can be reduced by 21% with hybrid-TES encoding. However, the consequence of this encoding method is the requirement in the symbol catalogue of four extra symbols. To offset the four symbol overhead by the 21% reduction in the transmission rate the complete symbol catalogue would have to comprise at least 20 symbols.

If the symbol catalogue is exactly 20 symbols, then no benefit is derived by using hybrid-TES for this particular

utterance. For utterances with a high source symbol generation rate or for utterances with highly emphasised fricatives, eg utterances BIRD and CBONLY, much higher reductions in transmission rates (70%, 51%) are possible. Disregarding 21% as the cost of overheads, 49% and 30% improvements in transmission rate are still possible when using hybrid-TES encoding.

If the symbol catalogue comprises more than 20 symbols then the benefits are obvious, even for the minimum reduction utterance.

If the catalogue comprises less than 20 symbols, then for the minimum reduction utterance it is non-beneficial to employ hybrid-TES encoding. However, benefits may be more apparent for utterances of the type BIRD or CBONLY.

Since in practice a fixed transmission rate will be used, the savings obtained for low epoch rate speech (eg utterance APPLE8) will not be significant, nor are they important. However, a communication system in general must be capable of handling utterances with particular emphasis on the unvoiced sounds. It is for such utterances that the hybrid-TES encoding scheme has been designed and that is where it is most beneficial.

Examination of the graphs in Figures 5.14-5.18 shows that fewer numbers of repeats are required with hybrid-TES. This is because with hybrid-TES the buffer size requirements are reduced and consequently the limit set for repeating is invoked less frequently, as compared with ordinary TES.

5.3 Quality Assessment

It is shown, objectively, in the previous section that for speech with highly emphasised unvoiced sounds reductions can be obtained in the transmission rate requirements (or buffer size and delay requirements), by using hybrid-TES encoding. This section presents a subjective assessment of the quality of the reconstructed speech.

Subjective appraisal of the speech quality may be made on the basis of intelligibility[51-53] and preference[54,55]. However, it is not the purpose of this investigation to measure the absolute quality of hybrid-TES, but merely to assess it in relation to the alternative proposal for reducing buffer size and delay, which is ordinary TES with the repeat strategy. Since both the repeat strategy and hybrid-TES are effective in the unvoiced regions, a relative assessment can be made by comparing the reconstructions obtained from the two techniques. Thus, a simple test procedure, where a group of listeners

is asked to show their preference of the two reconstruction techniques, is sufficient. In some respects, this simple procedure may be considered to follow some of the engineering practices recommended by the IEEE subcommittee[54] for the preference method of subjectively measuring the speech quality.

In the test procedure, comparison of similar effects can be ensured if for ordinary TES with the repeat strategy the buffer size is restricted to that which is required for hybrid-TES. In this way, using the buffer size requirement for hybrid-TES to limit the buffer size for ordinary TES, ensures that no repeats are required for hybrid-TES whereas repeats are necessarily invoked for ordinary TES. It is important to guarantee that no repeats are incurred in the hybrid-TES case, because the buffer simulation program does not have facilities for hybrid-TES reconstruction. In addition, to ensure that the only distortion effects considered are those due to the processing of unvoiced sounds , the voiced sounds are reconstructed from the original PCM samples.

5.3.1 Comparisons to be made

In chapter 4 two methods of hybrid-TES reconstruction were suggested - by bandpass filtering the noise source and by spectrally shaping it. Since it has not been determined which one of these methods is preferred, it is necessary

to compare both the methods individually with ordinary TES. In addition, for the more fricatively emphasised utterances, the same comparisons can be made between the reconstructions obtained at different transmission rates.

To further usefully employ the listening effort available, other desirable comparisons can be made. For example, a comparison between the spectrally shaped noise reconstruction and the bandpass filtered noise reconstruction. Although not directly relevant to this section, but useful later, a comparison between the spectrally shaped noise and the bandpass filtered noise reconstruction can be made at a different original speech bandwidth.

In summary, the following comparisons can be made:

(a) For all utterances. Comparison between spectrally shaped noise hybrid-TES (SSHT) and ordinary TES with repeat strategy (RST), both at 1000 symbols/sec.

ie: $SSHT_A$ - V - RST_A

(b) For all utterances. Comparison between bandpass filtered noise hybrid-TES (BHT) and ordinary TES with repeat strategy (RST), both at 1000 symbols/sec.

ie: BHT_B - V - RST_B

(c) For utterances CBONLY and BIRD. Comparison between spectrally shaped noise hybrid-TES at 1000 symbols/sec

and ordinary TES with repeat strategy at 1440 symbols/sec.

ie: $SSHT_C$ - V - RST_C

(d) For utterances CBONLY and BIRD. Comparison between bandpass filtered noise hybrid-TES at 1000 symbols/sec and ordinary TES with repeat strategy at 1440 symbols/sec.

ie: BHT_D - V - RST_D

(e) For utterance CBONLY. Comparison between spectrally shaped hybrid-TES at 1000 symbols/sec and ordinary TES with repeat strategy at 2000 symbols/sec.

ie: $SSHT_E$ - V - RST_E

(f) For all utterances. Comparison between spectrally shaped noise hybrid-TES and bandpass filtered noise hybrid-TES with the original speech at full bandwidth (10-4.5kHz).

ie: $SSHT_F$ - V - BHT_F

(g) For utterances CBONLY and BIRD. Comparison between spectrally shaped noise hybrid-TES and bandpass filtered noise hybrid-TES with the original speech bandlimited to 300-3.4kHz.

ie: $SSHT_G$ - V - BHT_G

5.3.2 Presentation to the listeners

The set of comparisons to be made for each of the utterances is shown in the upper half of Table 5.3. All these comparisons are repeated in reverse order to eliminate any local effects, as shown in the lower half of Table 5.3.

For the presentation to the listeners the comparisons were recorded in a random order, as shown in Figure 5.19. In order to ascertain the validity of the results, throughout the comparison set a control comparison was included after every seventh comparison. This control utterance took the form of a comparison between the unprocessed version of the utterance APPLE8 and a noise degraded version of it (SNR = 30dB).

The tests were conducted in a language laboratory which is capable of accommodating 20 listeners at a time. The listeners were presented with the instructions shown in Figure 5.20 and were asked to indicate the sentence of their preference in each of the comparison pairs that were played over the headphones, via the control centre.

5.3.3 Result analysis and discussion

Altogether twelve listeners volunteered to participate in the listening tests and of these five had previously been involved in speech research and may be classified as

APPLE8 (1)	CBONLY (2)	APPLE7 (3)	BIRD (4)	FEM1 (5)
$SSHT_A$-V-RST_A	$SSHT_A$-V-RST_A	$SSHT_A$-V-RST_A	$SSHT_A$-V-RST_A	$SSHT_A$-V-RST_A
BHT_B -V-RST_B	BHT_B -V-RST_B	BHT_B -V-RST_B	BHT_B -V-RST_B	BHT_B -V-RST_B
$SSHT_F$-V-BHT_F	$SSHT_C$-V-RST_C	$SSHT_F$-V-BHT_F	$SSHT_C$-V-RST_C	$SSHT_F$-V-BHT_F
PCM -V-NOISY	BHT_D -V-RST_D		BHT_D -V-RST_D	
	$SSHT_E$-V-RST_E		$SSHT_F$-V-BHT_F	
	$SSHT_F$-V-BHT_F		$SSHT_G$-V-BHT_G	
	$SSHT_G$-V-BHT_G			
RST_A -V-$SSHT_A$	RST_A -V-$SSHT_A$	RST_A -V-$SSHT_A$	RST_A -V-$SSHT_A$	RST_A -V-$SSHT_A$
RST_B -V-BHT_B	RST_B -V-BHT_B	RST_B -V-BHT_B	RST_B -V-BHT_B	RST_B -V-BHT_B
BHT_F -V-$SSHT_F$	RST_C -V-$SSHT_C$	BHT_F -V-$SSHT_F$	RST_C -V-$SSHT_C$	BHT_F -V-$SSHT_F$
NOISY-V-PCM	RST_D -V-BHT_D		RST_D -V-BHT_D	
	RST_E -V-$SSHT_E$		BHT_F -V-$SSHT_F$	
	BHT_F -V-$SSHT_F$		BHT_G -V-$SSHT_F$	
	BHT_G -V-$SSHT_G$			

TABLE 5.3

List of the subjective comparisons to be made

trained listeners. However, the results from the trained listeners are not treated differently to the results from the untrained listeners, because the effect of training was investigated separately by asking the listeners to repeat the tests. In this case eight listeners volunteered.

The first stage in the analysis process is to determine how the listeners performed with the control comparisons. Table 5.4 shows the results obtained from each of the listeners. Since listeners BL and MAMS were unable to give the correct response for the control comparisons their results are not considered in the remaining analysis. It seems that they have either misinterpreted the instructions or were unable to distinguish the fine differences.

The overall results obtained from the remaining listeners are summarised in Table 5.5, where the comparison pairs and their complements are grouped together. In the analysis of these results each of the utterances is examined individually and for each set of comparisons the results are analysed into three categories, as:

(i) Due to the first attempt listeners only.

(ii) Due to the first attempt and second attempt listeners, where the second attempt listeners are regarded as additional first attempt listeners.

Listeners	Decisions for control comparisons. A or B for preference-no preference				
T indicates trained listener	Pair (7)	Pair (14)	Pair (21)	Pair (28)	Pair (35)
DMC	-	B	B	B	A
T EW	A	B	B	B	A
T PCC	A	B	B	B	A
PJ	A	B	B	B	A
BL 1st	NI	A	-	B	A
BL 2nd	-	-	-	-	-
MAMS 1st	-	-	-	-	-
MAMS 2nd	-	-	A	-	-
TFD 1st	A	B	B	B	A
TFD 2nd	A	B	B	B	A
T PSC 1st	A	B	B	B	A
T PSC 2nd	A	B	B	B	A
T SL 1st	A	B	B	B	A
T SL 2nd	A	B	B	B	A
NH 1st	-	B	-	B	A
NH 2nd	A	B	B	B	A
AD 1st	-	B	B	B	A
AD 2nd	A	B	B	B	A
T RAK 1st	A	B	B	B	A
T RAK 2nd	A	B	B	B	A
Correct response :	A	B	B	B	A

TABLE 5.4

Subjective results for the control utterances

Listeners		APPLE8 (1&25)		APPLE8 (32&43)		APPLE8 (12&38)		CONTROL (7,14,21,28,35)					APPLE7 (41&50)		APPLE7 (2&15)		APPLE7 (26&45)		FEM1 (24&48)		FEM1 (19&30)		FEM1 (8&34)	
DMC	1	-	B	-	-	-	B	-	B	B	B	A	-	B	-	B	-	B	A	A	A	-	-	-
EW	1	-	-	-	-	-	-	A	B	B	B	A	-	-	-	-	-	-	-	-	-	-	-	-
PCC	1	B	-	B	-	-	-	A	B	B	B	A	-	-	B	-	A	-	A	A	-	A	-	-
PJ	1	B	B	B	A	-	A	A	B	B	B	A	-	B	A	-	-	-	-	-	B	A	-	A
TFD[1st]	1	B	B	A	-	B	-	A	B	B	B	A	B	A	A	A	-	-	A	-	-	-	-	-
TFD[2nd]	2	A	B	A	-	-	-	A	B	B	B	A	B	A	B	A	A	B	-	-	-	-	-	-
PSC[1st]	1	-	-	A	B	-	-	A	B	B	B	A	A	B	A	A	-	-	-	-	-	-	-	-
PSC[2nd]	2	-	B	B	-	-	-	A	B	B	B	A	A	B	-	B	-	-	-	-	-	-	-	-
SL[1st]	1	B	B	A	-	A	-	A	B	B	B	A	B	A	-	A	A	-	-	-	-	A	-	-
SL[2nd]	2	B	-	A	-	-	-	A	B	B	B	A	B	A	A	A	A	A	-	-	-	-	-	-
NH[1st]	1	A	-	-	A	-	-	-	B	-	B	A	-	-	-	A	-	-	A	B	-	A	-	-
NH[2nd]	2	-	-	-	B	-	-	A	B	B	B	A	-	-	B	A	-	-	A	B	-	A	-	A
AD[1st]	1	B	-	-	B	-	B	-	B	B	B	A	B	B	-	-	-	A	A	B	B	-	B	-
AD[2nd]	2	B	B	B	A	-	-	A	B	B	B	A	-	B	B	-	B	-	B	B	B	A	B	-
RAK[1st]	1	B	-	-	-	-	-	A	B	B	B	A	-	-	-	-	-	-	-	-	-	-	-	-
RAK[2nd]	2	-	-	-	-	-	-	A	B	B	B	A	-	-	B	-	-	-	-	-	-	B	-	-

TABLE 5.5

Summary of overall results from subjective listening tests

A or B for preference; - for no preference

/continued overleaf

Listeners	CBONLY														BIRD													
	(16&33)		(23&39)		(6&10)		(13&41)		(3&36)		(20&31)		(27&44)		(46&49)		(11&18)		(9&29)		(4&22)		(5&17)		(37&40)			
DMC	A	B	A	B	B	A	B	A	A	B	B	-	-	B	-	B	B	-	B	-	B	-	B	-	-	B		
EW T	A	B	A	B	B	A	B	A	A	-	A	-	A	B	A	B	B	A	B	A	A	B	-	-	-	A		
PCC T	A	B	A	B	B	A	B	B	A	B	-	A	B	B	B	A	B	A	B	A	A	A	A	A	A	B		
PJ	A	B	A	B	A	A	B	A	B	-	-	B	-	-	B	B	B	A	B	A	A	-	A	-	B	B		
TFD 1st	A	B	A	B	B	A	B	A	B	-	B	A	A	B	A	B	B	A	B	A	-	B	A	B	-	B		
TFD 2nd	A	B	A	B	B	A	B	A	-	A	B	A	A	B	A	B	B	A	B	A	A	B	B	A	A	B		
PSC 1st T	A	-	A	B	B	A	B	A	-	B	-	B	A	B	-	B	B	A	B	-	A	B	B	-	-	B		
PSC 2nd	A	B	A	B	B	A	B	A	-	A	B	B	A	B	-	B	B	A	B	-	A	B	A	-	-	B		
SL 1st T	A	B	A	B	B	A	B	A	-	A	-	A	A	B	A	B	B	A	B	A	A	B	A	B	-	B		
SL 2nd	A	B	A	B	B	A	B	A	B	-	B	-	A	B	A	B	B	A	B	A	A	B	A	B	B	B		
NH 1st	A	B	A	B	-	A	B	A	-	-	-	-	A	B	A	B	B	A	B	A	-	B	-	B	B	-		
NH 2nd	A	B	A	B	B	A	B	A	-	-	-	-	A	B	A	B	B	A	B	A	A	B	-	-	-	-		
AD 1st	A	B	A	B	B	A	A	B	A	A	B	B	A	B	-	B	-	B	B	B	B	-	-	B	A	-		
AD 2nd	A	B	A	B	B	A	B	B	A	B	B	B	B	B	A	A	B	A	B	-	-	B	B	B	B	-		
RAK 1st T	A	B	A	B	B	-	B	A	-	B	B	-	A	B	A	B	B	A	B	-	-	B	B	-	A	B		
RAK 2nd	A	B	A	E	B	A	A	A	-	B	B	-	A	B	A	B	B	A	B	A	A	B	B	A	B	B		

TABLE 5.5 (continued)

Summary of overall results from subjective listening tests

A or B for preference; - for no preference

(iii) Due to the second attempt listeners only, to investigate the effect of training.

The overall decision for each of these categories is made by averaging the decisions of all the listeners.

The analysis of results for utterance APPLE8 is summarised in Table 5.6. It is observed that for all the three categories there is no preference for either TES or hybrid-TES reconstruction. In addition, there is no preference for either bandpass noise reconstruction or for spectrally shaped noise reconstruction.

Table 5.7 and Table 5.8 summarise the analysis of results for utterance APPLE7 and FEM1 respectively. Except for the trained listeners category, there is no preference for either TES or hybrid-TES reconstruction. In the case of the trained listeners, preference is shown for hybrid-TES reconstruction with bandpass filtered noise. No preference is shown for the reconstruction of FEM1.

The analysis of results for the reconstruction of the utterance BIRD is shown in Table 5.9. In this case hybrid-TES is preferred by all categories of listeners, even when the transmission rate for TES is 1440 symbols/sec and for hybrid TES is 1000 symbols/sec. For the full bandwidth case, preference is shown for bandpass filtered

Utterance : APPLE8	Category (i) 1st attempt listeners			Category (ii) 1st & 2nd attempt listeners			Category (iii) 2nd attempt listeners only		
	Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A		
	NP	SSHT	RST	NP	SSHT	RST	NP	SSHT	RST
Individual (Pair (1)	3	1	6	6	2	8	3	1	2
decisions for (Pair(25)	6	4	0	9	7	0	3	3	0
Total	9	5	6	15	9	8	6	4	2
Overall decision	No preference			No preference			No preference		
	Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B		
	NP	BHT	RST	NP	BHT	RST	NP	BHT	RST
Individual (Pair(22)	5	3	2	7	5	4	2	2	2
decisions for (Pair(43)	6	2	2	10	3	3	4	1	1
Total	11	5	4	17	8	7	6	3	3
Overall decision	No preference			No preference			No preference		
	Comparison SSHT-V-BHT			Comparison SSHT-V-BHT			Comparison SSHT-V-BHT		
	NP	BHT	SSHT	NP	BHT	SSHT	NP	BHT	SSHT
Individual (Pair(12)	8	1	1	14	1	1	6	0	0
decisions for (Pair(38)	7	2	1	13	2	1	6	0	0
Total	15	3	2	27	3	2	12	0	0
Overall decision	No preference			No preference			No preference		

TABLE 5.6

Analysis of test results for the speech file APPLE8.SPH

Utterance : APPLE 7	Category (i) 1st attempt listeners			Category (ii) 1st & 2nd attempt listeners			Category (iii) 2nd attempt listeners only		
	Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A		
	NP	SSHT	RST	NP	SSHT	RST	NP	SSHT	RST
Individual (Pair(41)	6	3	1	9	5	2	3	2	1
decisions for (Pair(50)	4	2	4	6	6	4	2	2	2
Total	10	5	5	15	11	6	5	4	3
Overall decision	No preference			No preference			No preference		
	Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B		
	NP	BHT	RST	NP	BHT	RST	NP	BHT	RST
Individual (Pair (2)	6	1	3	7	5	4	1	4	1
decisions for (Pair(15)	5	4	1	7	7	2	2	3	1
Total	11	5	4	14	12	6	3	7	2
Overall decision	No preference			No preference			BHT preferred		
	Comparison SSHT-V-BHT			Comparison SSHT-V-BHT			Comparison SSHT-V-BHT		
	NP	BHT	SSHT	NP	BHT	SSHT	NP	BHT	SSHT
Individual (Pair(26)	8	0	2	11	1	4	3	1	2
decisions for (Pair(45)	8	1	1	12	2	2	4	1	1
Total	16	1	3	23	3	6	7	2	3
Overall decision	No preference			No preference			No preference		

TABLE 5.7

Analysis of test results for the speech file APPLE7.SPH

Utterance : FEM1	Category (i) 1st attempt listeners			Category (ii) 1st & 2nd attempt listeners			Category (iii) 2nd attempt listeners only		
	Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A		
	NP	SSHT	RST	NP	SSHT	RST	NP	SSHT	RST
Individual decisions for (Pair(24)	5	5	0	9	6	1	4	1	1
(Pair(48)	6	2	2	10	4	2	4	2	0
Total	11	7	2	19	10	3	8	3	1
Overall decision	No preference			No preference			No preference		
	Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B		
	NP	BHT	RST	NP	BHT	RST	NP	BHT	RST
Individual decisions for (Pair(19)	7	2	1	12	3	1	5	1	0
(Pair(30)	6	4	0	9	6	1	3	2	1
Total	13	6	1	21	9	2	8	3	1
Overall decision	No preference			No preference			No preference		
	Comparison SSHT-V-BHT			Comparison SSHT-V-BHT			Comparison SSHT-V-BHT		
	NP	BHT	SSHT	NP	BHT	SSHT	NP	BHT	SSHT
Individual decisions for (Pair (8)	9	1	0	14	2	0	5	1	0
(Pair(34)	9	1	0	14	2	0	5	1	0
Total	18	2	0	28	4	0	10	2	0
Overall decision	No preference			No preference			No preference		

TABLE 5.8

Analysis of test results for the speech file FEM1.SPH

UTTERANCE: BIRD	Category (i) 1st attempt listeners	Category (ii) 1st and 2nd attempt listeners	Category (iii) 2nd attempt listeners only
	Comparison $SSHT_A$-V-RST_A	Comparison $SSHT_A$-V-RST_A	Comparison $SSHT_A$-V-RST_A
	NP SSHT RST	NP SSHT RST	NP SSHT RST
Individual Pair (46)	3 5 2	4 10 2	1 5 0
decisions for Pair (49)	0 9 1	0 14 2	0 5 1
Total	3 14 3	4 24 4	1 10 1
Overall decision	Hybrid-TES preferred	Hybrid-TES preferred	Hybrid-TES preferred
	Comparison BHT_B-V-RST_B	Comparison BHT_B-V-RST_B	Comparison BHT_B-V-RST_B
	NP BHT RST	NP BHT RST	NP BHT RST
Individual Pair (11)	1 9 0	1 15 0	0 6 0
decisions for Pair (18)	1 6 1	1 14 1	0 6 0
Total	2 17 1	2 29 1	0 12 0
Overall decision	Hybrid-TES preferred	Hybrid-TES preferred	Hybrid-TES preferred
	Comparison $SSHT_C$-V-RST_C	Comparison $SSHT_C$-V-RST_C	Comparison $SSHT_C$-V-RST_C
	NP SSHT RST	NP SSHT RST	NP SSHT RST
Individual Pair (9)	0 10 0	0 16 0	0 6 0
decisions for Pair (29)	3 6 1	5 10 1	2 4 0
Total	3 16 1	5 26 1	2 10 0
Overall decision	Hybrid-TES preferred	Hybrid-TES preferred	Hybrid-TES preferred
	Comparison BHT_D-V-RST_D	Comparison BHT_D-V-RST_D	Comparison BHT_D-V-RST_D
	NP BHT RST	NP BHT RST	NP BHT RST
Individual Pair (4)	3 5 2	4 10 2	1 5 0
decisions for Pair (22)	3 6 1	3 12 1	0 6 0
Total	6 11 3	7 22 3	1 11 0
Overall decision	Hybrid-TES preferred	Hybrid-TES preferred	Hybrid-TES preferred
	Comparison $SSHT_F$-V-BHT_F	Comparison $SSHT_F$-V-BHT_F	Comparison $SSHT_F$-V-BHT_F
	NP BHT SSHT	NP BHT SSHT	NP BHT SSHT
Individual Pair (37)	5 3 2	7 4 5	2 1 3
decisions for Pair (40)	2 7 1	4 11 1	2 4 0
Total	7 10 3	11 15 6	4 5 3
Overall decision	Bandpass noise preferred	Bandpass noise preferred	Bandpass noise preferred
	Comparison $SSHT_G$-V-BHT_G	Comparison $SSHT_G$-V-BHT_G	Comparison $SSHT_G$-V-BHT_G
	NP BHT SSHT	NP BHT SSHT	NP BHT SSHT
Individual Pair (15)	3 3 4	4 6 6	1 3 2
decisions for Pair (17)	5 1 4	7 3 6	2 2 2
Total	8 4 8	11 9 12	3 5 4
Overall decision	No preference	Spectrally shaped noise preferred	Bandpass noise preferred

Table 5.9

Analysis of test results for the speech file BIRD.SPH

noise reconstruction, whereas for the limited bandwidth case the overall decision changes with the category of the listeners.

Table 5.10 shows the analysis for the utterance CBONLY. In this case too, hybrid-TES is preferred. Even when the comparison is between hybrid-TES at 1000 symbols/sec and ordinary TES at 2000 symbols/sec, preference is shown for hybrid-TES. For the full bandwidth case the trained listeners preferred the spectrally shaped noise reconstruction, although no preference is shown by the remaining listener categories. In addition, no preference is shown for the bandlimited case.

It can be concluded from this analysis that for speech with highly emphasised unvoiced sounds the hybrid-TES reconstruction strategy is preferred to ordinary TES with the repeat strategy. However, for speech with low unvoiced sound content there is no preference. In addition, it has been shown that generally there is no preference for the particular noise filter characteristics except that band-pass filtering is a simpler operation.

5.4 Probability Distributions

The real-zero interval probability distributions of the various speech utterances have already been presented. In this section a formal comparison is made of the

UTTERANCE: CBONLY		Category (i) 1st attempt listeners			Category (ii) 1st and 2nd attempt listeners			Category (iii) 2nd attempt listeners only		
		Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A			Comparison $SSHT_A$-V-RST_A		
		NP	SSHT	RST	NP	SSHT	RST	NP	SSHT	RST
Individual decisions for	Pair (16)	0	10	0	0	16	0	0	6	0
	Pair (33)	1	9	0	1	15	0	0	6	0
	Total	1	19	0	1	31	0	0	12	0
Overall decision		Hybrid-TES preferred			Hybrid-TES preferred			Hybrid-TES preferred		
		Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B			Comparison BHT_B-V-RST_B		
		NP	BHT	RST	NP	BHT	RST	NP	BHT	RST
Individual decisions for	Pair (23)	0	10	0	0	16	0	0	6	0
	Pair (39)	0	9	1	0	15	1	0	5	1
	Total	0	19	1	0	31	1	0	11	1
Overall decision		Hybrid-TES preferred			Hybrid-TES preferred			Hybrid-TES preferred		
		Comparison $SSHT_C$-V-RST_C			Comparison $SSHT_C$-V-RST_C			Comparison $SSHT_C$-V-RST_C		
		NP	SSHT	RST	NP	SSHT	RST	NP	SSHT	RST
Individual decisions for	Pair (6)	1	9	0	1	15	0	0	6	0
	Pair (10)	1	9	0	1	15	0	0	6	0
	Total	2	18	0	2	30	0	0	12	0
Overall decision		Hybrid-TES preferred			Hybrid-TES preferred			Hybrid-TES preferred		
		Comparison BHT_D-V-RST_D			Comparison BHT_D-V-RST_D			Comparison BHT_D-V-RST_D		
		NP	BHT	RST	NP	BHT	RST	NP	BHT	RST
Individual decisions for	Pair (47)	0	9	1	0	14	2	0	5	1
	Pair (47)	0	8	2	0	13	3	0	5	1
	Total	0	17	3	0	27	5	0	10	2
Overall decision		Hybrid-TES preferred			Hybrid-TES preferred			Hybrid-TES preferred		
		Comparison $SSHT_E$-V-RST_E			Comparison $SSHT_E$-V-RST_E			Comparison $SSHT_E$-V-RST_E		
		NP	SSHT	RST	NP	SSHT	RST	NP	SSHT	RST
Individual decisions for	Pair (27)	2	7	1	2	12	2	0	5	1
	Pair (44)	1	9	0	1	15	0	0	6	0
	Total	3	16	1	3	27	2	0	11	1
Overall decision		Hybrid-TES preferred			Hybrid-TES preferred			Hybrid-TES preferred		
		Comparison $SSHT_F$-V-BHT_F			Comparison $SSHT_F$-V-BHT_F			Comparison $SSHT_F$-V-BHT_F		
		NP	BHT	SSHT	NP	BHT	SSHT	NP	BHT	SSHT
Individual decisions for	Pair (20)	5	1	4	6	1	9	1	0	5
	Pair (31)	4	3	3	7	5	4	3	-	1
	Total	9	4	7	13	6	13	4	-	6
Overall decision		No preference			No preference			Spectrally shaped noise preferred		
		Comparison $SSHT_G$-V-BHT_G			Comparison $SSHT_G$-V-BHT_G			Comparison $SSHT_G$-V-BHT_G		
		NP	BHT	SSHT	NP	BHT	SSHT	NP	BHT	SSHT
Individual decisions for	Pair (3)	4	2	4	8	3	5	4	1	1
	Pair (36)	4	2	4	6	4	6	2	2	2
	Total	8	4	8	14	7	11	6	3	3
Overall decision		No preference			No preference			No preference		

Table 5.10

Analysis of test results for the speech file CBONLY.SPH

probability distributions at the various stages of processing, ie the original waveform, the intermediate processing state (IPS) and the reconstructed waveform.

The IPS distributions represent the probability of durations of the symbols stored in the transmitter buffer. The probability distribution of reconstructed speech is obtained by substituting filtered noise for regions of identified unvoiced sounds and by replacing the identified silence regions with zero amplitude samples.

When comparison is made between the original and the reconstructed distributions, the silence regions of the original waveform are also replaced by zero amplitude samples. In this way only the effects of processing the unvoiced regions are noticed.

Figures 5.21 - 5.25 show the probability distributions for the utterances of Appendix B. The IPS and the original waveform distributions are superimposed on the same axes so as to enable a direct comparison.

The distributions (Figures 5.23 and 5.24) for the utterances CBONLY and BIRD, show considerable reductions in the number of short duration epochs that have to be stored in the transmitter buffer, hence the saving in buffer size. Smaller savings are achieved in the

distributions (Figures 5.22 and 5.25) for the less fricative utterances (APPLE8 and APPLE7).

It is observed from the IPS distributions that hybrid-TES encoding does not completely eliminate the unvoiced peak. This is because of the presence of small amplitude, short duration epochs amongst the voiced epochs and also the requirement for the classification algorithm of a pre-set number of consecutive epochs having unvoiced parameters. The number of epochs remaining in the unvoiced peak after the IPS tend to be higher for noisy signals. This effect is illustrated by the distribution of utterance FEM1, which has a persistent hum in the background. Although the epochs due to this background appear in the unvoiced peak of the distribution, their characteristics are such that they are regarded as voiced epochs by the classification algorithm. As such, the unvoiced savings in the buffer size are rather small.

Confirmation of the effect of noise on the probability distribution is made by introducing noise into the utterance APPLE8. This utterance is chosen because the voiced/unvoiced peaks of its distribution are clearly distinguishable and the above effect can be easily illustrated. Two noisy files with signal to noise ratios of 20dB and 10dB are generated and their distributions measured. Examination of these distributions (Figures 5.26

and 5.27) shows that as the signal to noise ratio of the speech file decreases so also does the reduction in the unvoiced peak.

The distributions of the reconstructed utterances are shown in Figures 5.28 - 5.32. Again, the reconstructed and the original waveform distributions are superimposed on the same axes to enable direct comparisons. Because the reconstruction is achieved from a fixed number of patterns of noise, the reconstructed distributions do not exactly match the original distributions, but closely approximate them.

5.5 Signal Processing

This section presents the effects on hybrid-TES of signal processing, where the two processes to be considered are: the addition of noise, and the effects of filtering.

5.5.1 Addition of noise

In a real transmission system the corruption of a speech signal through background noise is inevitable, the amount and characteristics of the corruption depending on the particular background. It is important, therefore, to know the effect and limitations of hybrid-TES under these circumstances.

An investigation is conducted where the corruption

effect is modelled by a random noise generator. Since the processing and evaluation of results are repetitive and rather time consuming operations, only two speech files are used for the investigation. The speech files APPLE8 and CBONLY are selected on the basis of representing the extreme cases of low and high epoch rates.

A program, SPNOIS, is available which takes a speech file and gives three noisy versions each with a different signal to noise ratio. This is achieved by:

(i) generating a noise file and measuring its RMS value;

(ii) amending this RMS in accordance with the RMS of the speech file and the desired SNR;

(iii) adding the amended noise samples to the original speech samples to get the noisy signal;

(iv) amending the noisy speech samples to have the same peak value as the original speech file.

The resultant speech files are then processed using the hybrid-TES encoding algorithms and the effects noted.

Figures 5.33 and 5.34 present the effect on the classification algorithm of reducing the SNR. In Figure 5.33 the classification of the first twenty blocks of

speech file APPLE8 is shown for SNR of 30dB, 20dB and 10dB. The original speech waveform is superimposed for comparison purposes. It is observed that the differences in the classification occur mainly in the silence regions, for in these regions the energy levels of the noise epochs are high and often exceed the threshold. When this happens the classification algorithm merely decides between the voiced and unvoiced characteristics. The decision is made on the basis of the epoch durations. If the epoch durations are small (< 0.4ms) and occur consecutively, they are considered as unvoiced and if they are large (> 0.4ms) they are considered as voiced.

For the 30dB case, because the noise epoch energy levels sometimes exceed the threshold value, parts of the silence or inter-word regions are classified as either voiced or unvoiced. For the 20dB and 10dB case, silence is never detected, because the noise energy levels are always in excess of the threshold value.

Another effect of the reduction of SNR is the delay in detection of the true voiced regions. Again, because of the high energy levels of the noise epochs in the inter-word spaces, the first one or two epochs in the voiced sounds are often so low in energy that they too are regarded as noisy epochs.

Figure 5.34 shows the classification of the first twenty blocks of the speech file CBONLY for different SNRs. This file was processed exactly the same as in the previous case, with the exception that an attempt was made to locate the SNR at which the classification algorithm fails to function correctly. Comparison with the original waveform shows that the classification algorithm may be considered to fail at SNR = -5dB. However, audible analysis shows that the quality of the reconstructed speech is not unduly impaired at this SNR. In addition, the buffer does not suffer any detrimental effects, as a result of the false classification, since the number of symbols to be stored are reduced.

The reconstruction of hybrid-TES is a function of the classification algorithm. The effects of adding noise to speech are consequently relayed to the reconstruction process and are best analysed audibly.

The silence regions which have been classified as either voiced or unvoiced are reconstructed in exactly the same way as the true voiced or unvoiced regions. Because with lower SNRs more of the silence regions are classified as unvoiced, a "warbling" effect is observed in these regions of the reconstructed speech. This warbling effect is due to the random nature of the background noise model. The encoding algorithm uses the

average epoch rates of short segments of the unvoiced regions to differentiate between the different unvoiced symbols. With a random noise model the average epoch rate over short segments is continuously varying, although the average over longer periods is consistent with the bandwidth. Thus, the reconstructed regions, made up from a random combination of unvoiced symbols, results in a warbling effect. This effect is not apparent in the reconstruction of true unvoiced sounds because of the relatively invariant nature of the average epoch rate during these regions.

5.5.2 Filtering

Filtering of speech is a very important factor in the determination of the channel capacity requirements and the quality of speech - the lower the bandwidth, the lower the channel capacity and the poorer the speech quality. The actual constraints placed on the filtering operation depends on the particular application (for example, a telephone channel is bandlimited between 300Hz and 3400Hz). Since TES is potentially applicable to many speech communication systems, it is necessary to be aware of the effects of filtering on hybrid-TES. These effects can be broadly analysed in terms of classification and reconstruction.

Figures 5.35 and 5.36 show the effect on the

classification of bandlimiting the speech files APPLE8 and CBONLY, respectively. The bandwidth frequency limits imposed are 300Hz and 3400Hz. Only the blocks which show the greatest differences in the classification are illustrated here. These differences are mainly apparent in the silence regions and do not represent any serious discrepancies.

An analytic method of investigating the effect of filtering on the reconstruction quality is through determination of the epoch count. Figure 5.37 shows the effect on the epoch count of low-pass filtering the speech with different cut-off frequencies of the 2nd order Butterworth digital filter. It is observed that the lower the cut-off frequency, the lower the epoch count; and that the reduction is approximately linear. In addition, increasing the filter order causes the epoch count to be further reduced.

Figure 5.38 shows the effect on the epoch count of high-pass filtering with different cut-off frequencies. As the cut-off frequency of the filter is increased, the higher voiced frequencies become more emphasised and thus cause an increase in the epoch count.

Due to the complementary behaviour of the high-pass and low-pass filtering operations the overall effect of

bandpass filtering the speech depends on the respective cut-off frequencies of the bandpass filter. One observation which can be made is that, unless specifically arranged by the choice of the cut-off frequencies, the average epoch rate of bandlimited speech would be different from the full bandwidth case. Confirmation of this is made by bandpass filtering the utterances CBONLY and APPLE8 between 300Hz and 3400Hz. The epoch count changes from 4339 to 4747 for the utterance CBONLY and 1929 to 2655 for the utterance APPLE8.

Since hybrid-TES is effective in the unvoiced regions and these are high frequency sounds, it is the upper cut-off frequency of the bandpass filter which affects hybrid-TES reconstruction. The variation of the upper cut-off frequency changes the average epoch rate during the unvoiced sounds. This variation in the epoch rate would present no problem to the bandpass filtered noise reconstruction technique. However, for the spectrally shaped noise reconstruction, if the variation in the epoch rate exceeds that range of the sounds which each of the noise sources is designed to cover, distortion may result. Experiments conducted in the quality section, however, have shown that in the reconstruction of speech bandlimited between 300-3400Hz, generally, there was no preference between the bandpass filtered noise reconstruction and the spectrally shaped noise reconstruction.

On the other hand, from the point of view of being able to cater for all bandwidth cases, the bandpass filtered noise reconstruction would appear preferable.

CHAPTER 6

HYBRID-TES ENCODING OF SPEECH

6.1 Introduction

Having investigated the feasibility aspect of hybrid-TES a formal presentation can now be made of its encoding and reconstruction strategies, bearing in mind the real-time implications.

A typical hybrid-TES transmission system is shown, in block diagram form, in Figure 6.1. It is called hybrid-TES because it uses different encoding methods for the different regions of speech sound. It differs from ordinary TES in that the speech waveform is first classified into three different categories: voiced, unvoiced and silence. Each of the categories is then separately encoded. The voiced regions are encoded as normal TES; and the unvoiced and silence regions are encoded in 5ms segments. Those unvoiced or silence regions which have less than 5ms durations are encoded as single voiced epochs, with appropriate amplitudes and durations. At the receiver, the unvoiced symbols are used to inject filtered noise for the required durations.

6.2 Encoding

The epoch parameters required for ordinary TES encoding are duration, ED; amplitude, EA; and shape, ES.

Hybrid-TES makes use of precisely these same parameters. The classification algorithm uses the amplitude and duration information of successive epochs for determining the class, EC, of each individual epoch, ie the determination of whether an epoch is due to voiced sounds, unvoiced sounds or silence. Depending on the class, one of three procedures takes place and these are described below. Figure 6.2 shows the general flow diagram for the complete encoding operation.

6.2.1 Unvoiced epochs

When the epoch class is identified as unvoiced, the encoding procedure is as shown in the flow diagram of Figure 6.3.

Since the unvoiced and silence regions of speech are to be encoded as segments of 5ms duration, there will be occasions when these identified regions are not exact multiples of 5ms. For this reason the end portion (less than 5ms) of the identified regions is encoded as a simple shaped, voiced, epoch having the appropriate duration and amplitude. Thus the first stage in the flow diagram is to determine if the class of epochs has just changed from silence to unvoiced. If this is the case, then the stored information (SED, SEA), which is the end portion of the silence region before change of class to unvoiced, has to be encoded first. If SED = 0, there is no silence information to be encoded. If SED $\neq$ 0, then a voiced symbol

having a duration, SED, amplitude SEA, and a simple shape is transferred to the buffer. SED and SEA are reset.

The next stage is to process the unvoiced epochs. The duration, ED, of each of the epochs is added to a running total, UED. The amplitude, EA, is used to determine a running average amplitude, UEA. When UED > K_1, an unvoiced symbol is transferred to the buffer. The decision as to which one of the four unvoiced symbols should be transferred is made on the basis of the number of epochs, NOE, required to make up the segment of duration K_1. For example, symbol 1 is transferred when NOE > K_2 and symbol 3 when NOE > K_4. NOE and UEA are then reset to 0 and EA respectively. UED is again compared with K_1 as a precautionary measure in case an epoch larger than duration K_1 has been identified as an unvoiced epoch. The constants K_1, K_2, K_3 and K_4 have been determined by previous experiment as 5ms, 2800, 4800 and 6800 respectively.

6.2.2 Silence epochs

The flow diagram for the processing of silence epochs is shown in Figure 6.4. As in the previous case, it is necessary first to determine and encode any stored information about the unvoiced region which was remaining before the change of class to silence. It is encoded as a voiced symbol having duration UED and amplitude UEA.

The silence epochs are processed in a similar manner to the unvoiced epochs, with the exception that there is only one symbol which is transferred to the buffer. Since some form of amplitude information is to be transmitted, the silence symbol can be regarded as one of the unvoiced symbols (eg symbol 4) but with zero amplitude. In this way the transmission of silence does not incur any addition to the number of symbols in the catalogue.

In the flow diagram, SED is the running total of the silence duration. When SED > K_1 a symbol of duration K_1 with zero amplitude is transferred to the buffer. SED is reduced by K_1 and once again compared with K_1. This comparison is necessary because epochs of very long duration have been known to exist in the silence regions. For example, if SED < 5ms and ED of the next epoch is 10ms, then after transferring a 5ms symbol to the buffer SED > 5ms. Thus, before returning to control ED of the following epoch another 5ms silence symbol must be transferred to the buffer to ensure that SED < 5ms. Although SEA is not directly used in encoding the actual silence symbols, it is used in encoding the end portions of the silence regions, ie the portions of the silence regions which have to be represented by voiced symbols.

6.2.3 Voiced epochs

For each voiced epoch a symbol is transferred to the

buffer, having a duration, VED; amplitude VEA; and shape, VES. In this case VED, VEA and VES are the same as ED, EA and ES respectively. On occasions, epochs with very large durations may be encountered. Since it is impractical to have encoding symbols for all possible durations, a limit has to be placed on the maximum encodable duration.

It has been previously determined that the maximum unvoiced or silence segments to be encoded are 5ms. Thus, the end portions of the unvoiced or silence regions can be up to 5ms in duration and have to be encoded using a voiced symbol. Therefore, voiced symbols must cover a duration range of at least 5ms. When a voiced epoch with ED > 5ms is encountered, each 5ms segment of the epoch is encoded by a voiced symbol of VED = 5ms, VEA = EA and a simple shape. The remaining end portion of the voiced epoch is encoded by another voiced symbol. Figure 6.6 illustrates the concept.

Since epochs exceeding 5ms in duration are rare in the voiced regions, this strategy of encoding is unlikely to cause any noticeable distortion.

The flow diagram for the processing of voiced epochs is shown in Figure 6.5. As in both the previous cases, it is necessary to determine and encode any unvoiced or silence regions remaining before change of class to voiced.

6.3 Decoding and Reconstruction

The received symbols are stored in the buffer and decoded one at a time. For TES, stored segment forms are reproduced in sequence at the correct duration in accordance with the specification corresponding to the received symbol. For hybrid-TES the reconstruction procedure is somewhat different when the unvoiced symbols are received. In such cases reconstruction is achieved by inserting filtered noise for a fixed duration.

It is important, therefore, in the decoding procedure of hybrid-TES, to discriminate between the two types of symbol, ie voiced and unvoiced. Depending upon the type of symbol a different reconstruction procedure is adopted.

6.3.1 Unvoiced symbols

It has been shown previously that four different noise filter characteristics are required for adequate reconstruction of the unvoiced sounds. Information about the frequency characteristics of the filters is conveyed by the four different unvoiced symbols. Thus, discrimination has to be made between the four different symbols to allow the appropriate noise filter to be selected.

Figure 6.7 shows the flow diagram for the reconstruction from unvoiced symbols. After symbol discrimination the noise source is appropriately filtered and the

resultant noise, with appropriate amplitude, is injected onto the line for a duration of 5ms.

Discrimination between symbol 4 representing a segment of the unvoiced or the silence region is made on the basis of the amplitude information received. If the amplitude information is zero, then symbol 4 represents a silence segment and reconstruction is made by reducing the amplitude of the noise source to zero. If the amplitude information is non-zero, then the noise amplitude is set to the appropriate value.

6.3.2 Voiced symbols

The voiced regions are reconstructed in exactly the same way as ordinary TES. Each of the symbols is used to define the duration, amplitude and shape of the stored segment form which is to be used for reconstruction.

The actual nature of the stored segment forms depends on the particular reconstruction strategy used; for example, sinusoidal reconstruction, parabolic reconstruction, quadratic reconstruction or a combination of these.

6.4 Real-Time Considerations

Hybrid-TES encoding obviously presents an additional processing load. As a result, the transmitter and

receiver complexity and processing times are increased. The real-time implications of this additional processing load, due to the encoding and reconstruction algorithms, are now considered.

6.4.1 System complexity

The system complexity is increased as a result of the extra memory space required.

At the transmitter, there is a need for storing the program code for both the classification and the encoding algorithms. Additional memory space is required for the processing of measured parameters.

The exact extent of the complexity can only be established if the design is made with a particular real-time system in mind. Since, due to the unavailability of time, this is not possible here, it is left as a recommendation for future work. On the other hand, an approximation to the order of the additional memory space required can be provided by considering the system flow diagrams in terms of simple source programs for the Motorola M6800 microprocessor.

Figure 6.8 shows the source program for the classification algorithm and Figure 6.9 shows the program for the encoding algorithm. These programs are not exact replicas

of the flow diagrams, but merely simple interpretations to allow an estimation to be made. It is observed from this simple interpretation of hybrid-TES that approximately 140 bytes are required for the classification algorithm, and approximately 200 bytes for the encoding algorithm. Inclusion of such features as the setting and resetting of flags, to indicate the availability of measured parameters, and the processing of these parameters, could bring the total requirement to about 500 bytes of additional memory at the transmitter.

At the receiver hybrid-TES requires that a filtered noise source be available for reconstruction purposes. This can be provided in one of two ways. Either four previously filtered noise files can be generated and stored in memory and used when required, or a physical noise source and a bank of four filters can be made available so that the selected filter is inserted for reconstructing the unvoiced sounds. Both methods have their drawbacks. In the former, to ensure a sufficiently random noise signal, a large memory area is required for each of the files - equivalent to about six blocks of speech. For the four files, 24 blocks or 24 x 256 = 6144 memory cells, ie a 6K byte memory unit is required. In the latter case, the interswitching of filters would give rise to unwanted transients. As for the reconstruction source program, approximately 50 bytes should suffice.

6.4.2 Processing times

Each epoch is processed individually and in three stages: parameter measurement, determination of class, and encoding. As soon as the parameter measurements become available, the epoch class is determined. In the classification algorithm, the most time consuming process is the determination of the average peak amplitude, which involves one multiplication and one division operation. However, in the source program an effort to reduce the processing times has been made by using an approximate method to determine the running average. This involves the use of the divide by 2^n (or shift) statements. Moreover, the source program has been written to enable a direct evaluation of the approximate processing times.

Consecutive operation of the classification and the encoding algorithms requires approximately 150 machine cycles of additional processing time. The total, including the setting and resetting of flags, would certainly be less than 200 machine cycles. Providing the epoch resolution is kept to below 200 machine cycles, the processing for hybrid-TES can be adequately performed. Figure 6.10 shows a typical timing chart for hybrid-TES.

CHAPTER 7

CONCLUSIONS

The requirement of large buffer size and consequent delay for TES during the unvoiced sounds has been confirmed by measuring the long-term real-zero probability distribution of speech (RZPDS).

The measurements (Chapter 2) were conducted in real-time and by using a microprocessor to perform the timing and counting functions. In the distribution two peaks were observed: one due to voiced sounds in the range 1ms - 2ms; and the other, sometimes larger one, due to unvoiced sounds in the range 0 - 500μs.

A further insight into the distribution pattern was provided by measuring the actual variation of epoch rate against time. Here, it was observed that the epoch rate varies from about 500 epochs/sec in the voiced regions to about 6000 epochs/sec in the unvoiced region. In addition, it was observed that the unvoiced sounds result in the bunching of epochs having similar characteristics. Consequently, a hybrid-TES means of encoding was proposed, whereby the voiced regions are to be encoded using TES parameters and the unvoiced regions by special symbols, indicating reconstruction with spectrally shaped random

noise. In this way, not only was the overall transmission rate expected to be reduced, but also the buffer size and consequent delay.

An investigation of the feasibility of this encoding strategy was conducted in three states, as:

(1) Identification of the unvoiced regions of the speech waveform.

(2) Substitution of the identified regions by spectrally shaped random noise and the determination of the encoding and reconstruction parameters.

(3) Performance analysis and comparison between ordinary TES and hybrid-TES.

In the identification process (Chapter 3) a classification algorithm was developed which successfully separates the voiced, unvoiced and silence regions of speech. Here, success was defined as no momentary changes of class whilst in the voiced region, although permitting some sensitivity in the silence and unvoiced regions.

The algorithm operates on epochs and the class of each epoch is decided on the basis of the epoch amplitude and duration and the trend of previous such measurements. Discrimination between silence and non-silence is made

by comparing the epoch amplitude with a threshold level set at 38dB below the signal peak. Discrimination between voiced and unvoiced is made by comparing the epoch duration with a reference set at 400μs. The transition regions between the voiced and the unvoiced are classified as voiced. This is achieved by transferring from unvoiced to voiced upon detection of two consecutive epochs having durations greater than 400μs, and transferring from voiced to unvoiced upon detection of 30 successive epochs having durations less than 400μs. The transitions between silence and unvoiced are classified as unvoiced. This is achieved by transferring from unvoiced to silence , when the average epoch amplitude of the previous 30 epochs decreases to less than the threshold of 38dB below the signal peak; and transferring from silence to unvoiced upon detection of two successive epochs having amplitude in excess of the threshold.

Replacement of the identified silence regions of the speech files with zero amplitude gave epoch reductions of between 10% and 20%.

In the substitution process (Chapter 4) identification of the different unvoiced sounds was attempted on the basis of the average epoch rate. It has been shown that although it is possible to generate each of the unvoiced sounds from spectrally shaped random noise, identification of

the individual sounds is not possible by epoch rate alone. On the other hand, it has been shown that adequate reconstruction can be obtained by using only four different noise sources. This has been possible because some of the important sound identification cues reside in the transition regions, and the correct coding of these regions is ensured by the classification algorithm. Thus, the noise sources only provide an approximation to the spectra of the sounds, which has been shown to be separable into three categories. These categories are low pass, band pass and high pass. Reconstruction of the sounds was achieved in two ways:

(i) By bandpass filtering the four noise sources with 2nd order Butterworth filters in the frequency ranges: 0.8-1.2kHz; 1.7-2.1kHz; 2.7-3.1kHz; 3.7-4.1kHz.

(ii) By spectrally shaping the four noise sources as: 2nd order Butterworth low-pass at 1.5kHz; 2nd order Chebyshev low-pass with 10dB ripple and cut-off at 3.3kHz; 2nd order Butterworth high pass at 3.0kHz; 3rd order Butterworth high pass at 4.0kHz.

It was confirmed by subjective listening tests that spectral shaping the noise sources gives sharper reconstruction.

A segmenting algorithm has been developed which enables the addition to the catalogue of unvoiced symbols to be kept at a minimum. In this way the unvoiced sounds are transmitted in 5ms segments and only four additional symbols are required in the catalogue. Moreover, the same segmenting algorithm has been proposed for efficiently transmitting the silence regions of speech, with the exception that the amplitude information accompanying the unvoiced symbol, in this case, is zero. In this way no additional symbols are required for transmitting silence. For the case where the unvoiced or silence regions are not exact multiples of 5ms, the end portion is encoded as a voiced symbol.

In the performance analysis (Chapter 5) it has been shown that for low epoch rate utterances (eg APPLE8), there is no difference between using TES and hybrid-TES; whereas for utterances with a high epoch rate (eg CBONLY), the transmission rate can be reduced considerably (49%) by using hybrid-TES. Alternatively, for a fixed transmission rate, the buffer size and consequent delay can be reduced by about 90%. For a system with a fixed transmission rate, buffer size and delay, the reductions obtained by using hybrid-TES may be considered in terms of the number of repeats, as compared with ordinary TES with repeat strategy.

The effect of the reduction of repeats was investigated through subjective listening tests. It has been shown that for the high epoch rate utterances at the same transmission rate, buffer size and delay, the listeners preferred hybrid-TES reconstruction to ordinary TES with repeat strategy. However, for low epoch rate utterances, no preference was shown.

Examination of the real-zero probability distributions at the various stages of processing has shown that for high epoch rate utterances the unvoiced peak is considerably reduced in the probability distribution of the symbols to be stored in the transmitter buffer. Pattern of the real-zero interval probability distribution of the reconstructed speech has been shown to closely match the pattern of the original speech distribution.

The effect on hybrid-TES of transmitting noisy signals was investigated by adding noise to the speech files at various signal to noise ratios. It was observed that for reduced signal to noise ratios silence was never detected. Instead, the classification algorithm simply decided between voiced and unvoiced speech. Because some of the voiced regions were classified as unvoiced, the classification algorithm was considered to fail at SNR = -5dB.

Reconstruction of the reduced SNR speech gave rise to a "warbling" effect in the inter-word spaces. This resulted from the reconstruction of random noise by the spectrally shaped noise used for reconstructing the unvoiced sounds.

The effect on the classification and reconstruction of bandpass filtering the original speech between 300-3400Hz was investigated and shown to be negligible.

Consideration of the real-time implementation of hybrid-TES has shown that the system complexity and processing times increase. The complexity increases as a result of the requirement of additional memory space for processing the measured parameters and storing the source program. Approximately 500 bytes of additional memory would be required at the transmitter and approximately 6000 bytes at the receiver. The total execution time of the classification and encoding algorithms approaches 200 machine cycles.

Although the speech material available for processing was limited, it does represent a fair cross-section of the type of utterances which may be encountered in practice. Therefore, the above conclusions are valid and should be regarded as general.

RECOMMENDATIONS FOR FUTURE WORK

An investigation of hybrid-TES has been presented and it is shown to be a feasible solution to the problem with ordinary TES of requiring a large buffer size and delay whilst transmitting unvoiced sounds. However, due to the unavailability of time some areas have not been fully investigated.

This section gives an indication of those areas which require further investigation.

(i) The classification algorithm. In the determination of reference parameters for this algorithm (eg the threshold and the epoch duration), the introduction of distortion was avoided as much as possible. If these parameters are set more critically, whereby the introduction of a controlled amount of distortion is allowed, then much larger reductions can be obtained in either the transmission rate, or buffer size and delay. In this way, more areas of the speech waveform would be identified as silence or unvoiced, and therefore fewer symbols would be stored in the buffer. In addition, when speech is coded using the actual TES symbols, the distortion introduced due to the classification algorithms may well be indistinguishable from the distortion introduced

as a result of the reconstruction from TES symbols.

The reference threshold value has been related to the level of speech signal via its peak amplitude. Since the peak amplitude is likely to be different for different speakers, there is a requirement for developing some form of a peak sensing algorithm. The threshold level can then be related to the speech signal by setting it at 38dB below the value determined by the peak sensing algorithm.

(ii) The reconstruction process. Because the intended application of hybrid-TES was bandlimited speech, identification of the different unvoiced sounds was attempted on the basis of the average epoch rate of normal speech, even though it is envisaged that pre-emphasis of speech may be used for TES. Since pre-emphasis of bandlimited speech tends to equalise the epoch rates of voiced and unvoiced sounds, it would be difficult to distinguish between them. However, for broadband speech, unvoiced sound discrimination is more convenient through pre-emphasis of the speech signal, as stated by Ito and Donaldson[43]. Thus, it is likely that for broadband speech pre-emphasis is likely to enable better quality reconstruction, although a larger number of

unvoiced symbols would be required.

(iii) Real-time implementation. It has been shown that implementation of the classification and encoding algorithms is possible on the Motorola M6800 microprocessor, if the epoch resolution is limited to about 200 machine cycles. With a 1MHz clock frequency, this is about 200μs and is considered to be too coarse for TES. However, with 2MHz 8-bit microprocessors, the resolution can be improved to about 100μs, which is certainly feasible for 10kHz sampled speech.

In the estimation of execution times, the setting and resetting of flags, which are required to indicate the availability of epoch data, were generously allocated 50 machine cycles. It is expected that actual implementation of these functions should enable faster execution times. If, however, this is not possible, implementation has to be considered on alternative processors, which should enable faster execution as a result of their improved architectures and faster clock frequencies.

Work in this area is not only useful for confirming the implementation in real-time, but also in extending the investigation of areas (i) and

(ii). That is, the effect on speech quality of varying any of the reference parameters can be observed directly in real-time. In addition much more speech can be processed, hence enabling more general conclusions. It is thus recommended that, since the previous two areas of investigation are refinements to the basic algorithms, priority be given to the real-time implementation.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the following, who have enabled the completion of this thesis.

(i) Professor J F Eastham, Head of School of Electrical Engineering at the University of Bath, for providing the facilities.

(ii) Mr J D Martin, Project Supervisor, for his constant encouragement and assistance in the difficult times encountered during the completion of the project.

(iii) The Science & Engineering Research Council for their financial assistance.

(iv) The speech research group at the University of Bath, which includes: Mr R D Hughes, Mr P S Cooper, Dr P C Ching, Mr A Seneviratne and Mr S Longshaw, for their general assistance and understanding.

(v) Mrs Diane Milton for her time and effort in carefully typing the thesis.

REFERENCES

1. Martin, J.: "Future Developments in Telecommunications", Prentice-Hall, London, 1977.

2. Flanagan, J.L.: "Speech analysis, synthesis and perception", Springer-Verlag, Berlin-Heidelberg, 1965.

3. Crowley, T.H., Harris, G.G., Miller, S.E., Pierce, J.R. and Runyan, J.P.: "Modern Communications", Chapter 6, pp.110-128, Columbia University Press, 1962.

4. Shannon, C.E.: "Communication in the presence of noise", Proc. Inst. Radio Engineers, 37, pp.10-21, January 1949.

5. Williams, J.S.: "Some bandwidth compression systems for speech transmission", The Radio and Electronic Engineer, pp.459-471, June 1946.

6. Gabor, D.: "New possibilities in speech transmission", J. Inst. Elect. Engrs., 94, Part III, pp.369-90, November 1947.

7. Fairbanks, G., Everitt, W.C. and Jaegar, J.P.: "Method for time or frequency compression - expansion of speech", I.R.E. Convention Record, Part 8, pp.120-124, 1953.

8. Bullington, K. and Fraser, J.M.: "Engineering aspects of t.a.s.i.", Bell Syst. Tech. Journal, 38, pp.353-64, March 1959.

9. Miller, G.A. and Licklider, J.C.R.: "Intelligibility of interrupted speech", J. Acoust. Soc. Amer., 22, pp.167-173, March 1980.

10. Wathen-Dunn, W. and Lipke, D.W.: "On the power gained by clipping speech in the audio band", J. Acoust. Soc. Amer., 30, pp.36-40, January 1958.

11. Licklider, J.C.R.: "Effects of amplitude distortion upon the intelligibility of speech", J. Acoust. Soc. Amer., 18, Part II, pp.429-435, Oct.1946.

12. Kelly, L.C.: "Speech and vocoders", The Radio and Electronic Engineer, pp.73-82, August 1970.

13. Holmes, J.N.: "The JSRU Channel Vocoder", Proc.IEEE, Vol.127, Part F, No.1, Feb.1980.

14. Schroeder, M.R.: "Vocoders: Analysis and synthesis of speech", Proc. IEEE, Vol.54, pp.720-734, May 1966.

15. Betts, J.A.: "Signal processing, modulation and noise", ELBS, Bristol, 1975.

16. Taub, H. and Schilling, D.L.: "Principles of communication systems", McGraw-Hill, 1971.

17. Abate, J.E.: "Linear and adaptive delta modulation", Proc. IEEE, Vol.55, pp.298-308, March 1967.

18. Jayant, N.S.: "Digital coding of speech waveforms", Proc. IEEE, Vol.62, pp.611-632, May 1974.

19. Atal, B.S. and Hanauer, S.L.: "Speech analysis and synthesis by linear prediction of the speech wave", J. Acoust. Soc. Amer., Vol.501, Part 2, pp.637-655, August 1971.

20. Atal, B.S. and Schroeder, M.R.: "Adaptive predictive coding of speech signals", Bell System Tech. J., Vol.49, 1970.

21. King, R.A. and Gosling, W.: "Time encoded speech", Electronics Letters, Vol.14, No.15, 29 July 1978.

22. Gosling, W. and King. R.A.: "Time encoded speech - an approach to digital voice transmission", Telephony, April 14, 1980.

23. King, R.A. and Gosling, W.: "Time encoded speech", Internal Research Report, University of Bath.

24. Gosling, W.: "Time encoded speech: The logic of the Bath program", presentation to the IEE Conference at Shrivenham, 1978.

25. Davenport, W.B.: "An experimental study of speech wave probability distributions", J. Acoust. Soc. Amer., Vol.24, No.4, pp.390-399, July 1952.

26. Singh, A.: "Measurement of real-zero probability distribution of speech using a microprocessor", Internal Research Report, School of Electrical Engineering, University of Bath, 1979.

27. Sturges, H.A.: "The choice of a class interval", J. Am. Stat. Assoc., Vol.21, pp.65-66, 1926.

28. Flanagan, J.L.: "Voices of men and machines", J. Acoust. Soc. Amer., Vol.51, No.5, Part I, 1972.

29. Fant, G.: "The acoustics of speech", Proc. of the 3rd International Congress on Acoustics, pp.188-201, 1959.

30. Fletcher, H.: "Speech and hearing in communications", Van Nostrand, 1953.

31. Holmes, J.N.: "Speech synthesis", Mills and Boon Ltd, London, 1972.

32. Gold, B.: "Note on buzz-hiss detection", J. Acoust. Soc. Amer., Vol.36, pp.1659-1661, 1964.

33. Noll, A.M.: "Cepstrum pitch determination", J. Acoust. Soc. Amer., Vol.41, pp.293-309, Feb., 1967.

34. Sondhi, M.M.: "New methods of pitch extraction", IEEE Trans. on Audio Electroacoust., Vol.AU-16, pp.262-266, June 1968.

35. Markel, J.D.: "The SIFT algorithm for fundamental frequency estimation", IEEE Trans. on Audio Electroacoust., Vol.AU-20, pp.367-377, Dec.1972.

36. Atal, B.S. and Rabiner, L.R.: "Apattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol.ASSP-24, No.3, June 1976.

37. Siegal, L.J.: "A procedure for using Pattern classification techniques to obtain a voiced/unvoiced classifier", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol.ASSP-27, No.1, Feb.1979.

38. Knorr, S.G.: "Reliable voiced/unvoiced decision", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol. ASSP-27, No.3, June 1979.

39. Baker, J.M.: "Time-domain analysis and segmentation of connected speech", Speech Communication, Proc. of the Speech Communication Seminar, Stockholm, Vol.3, pp.369-383, April 1-3, 1974.

40. Al-Doubooni, M.M.Z.: "Speech encoding for low data rate transmission", PhD Thesis, University of Bath, 1981.

41. Halle, M., Hughes, G.W. and Radley, J.P.A.: "Acoustic properties of stop consonants", J. Acoust. Soc. Amer., Vol.29, No.1, Jan.1957.

42. Hughes, G.W. and Halle, M.: "Spectral properties of fricative consonants", J. Acoust. Soc. Amer., Vol.28, No.2, March 1956.

43. Ito, M.R. and Donaldson, R.W.: "Zero-crossing measurements for analysis and recognition of speech sounds", IEEE Trans. on Audio Electroacoust., Vol.AU-19, pp.235-292, 1971.

44. Heinz, J.M. and Stevens, K.N.: "On the properties of voiceless fricative consonants", J. Acoust. Soc. Amer., Vol.33, No.5, May 1961.

45. Rice, S.O.: "Statistical properties of random noise", Bell Sys. Tech. J., Vol.24, pp.46-159, Jan.1945 (Part III of Mathematical Analysis of Random Noise by S.O. Rice, July 1944).

46. Potter, R.K., Kopp, G.A. and Green, H.C.: "Visible speech", D. Van Nostrand Company Inc., Princeton, New Jersey, 1947.

47. Turner, L.F., Frangoulis, E. and Alcaim, A.: "Some considerations relating to the performance of variable-information-rate-source to constant-transmission-rate schemes of data compression", Computers and Digital Techniques, Vol.2, No.3, June 1979.

48. Mason, D.C. and Balston, D.M.: "Relationship between system delay and transmission rate in time encoded speech", Electronics Letters, Vol.16, No.4, 14 Feb.1980.

49. Seneviratne, A.: "The force repeat strategy of buffer management for TES", Internal Research Report, School of Electrical Engineering, University of Bath, Nov.1981.

50. Martin, J.D. and Seneviratne, A.: "A simulation study of buffer overflow in a TES system", Internal Research Report, School of Electrical Engineering, University of Bath, July 1981.

51. Agrawal, A.: "An on-line speech intelligibility measurement system", IEEE Trans. on Acoust. Speech and Sig. Proc., Vol.ASSP-22, No.3, June 1974.

52. Wong, D.Y. and Markel, J.D.: "An intelligibility evaluation of several linear prediction vocoder modifications", IEEE Trans. of Acoust. Speech and Sig. Proc., Vol.ASSP-26, No.5, Oct.1978.

53. Voiers, W.D.: "The present state of digital vocoding techniques - a diagnostic evaluation", IEEE Trans. on Electroacoust., Vol. AU-16, pp.275-279, June 1968.

54. IEEE recommended practice for speech quality measurements, IEEE Trans. on Audio Electroacoust., Vol.AU-17, pp.227-246, Sept.1969.

55. Rothauser, E.H., Urbanek, G.E. and Pachl, W.P.: "A comparison of preference measurement methods", J. Acoust. Soc. Amer., Vol.49, No.4, Part 2, 1971.

56. Motorola "M6800 Microprocessor Applications Manual", Motorola Inc., Switzerland, 1975.

APPENDIX A

SOFTWARE DESIGN FOR THE MICROPROCESSOR METHOD OF MEASURING THE RZPDS

In Chapter 2 it was shown how a microprocessor can be used to perform the timing and counting functions in the measurement of the real-zero probability distribution of speech. Here, the actual microprocessor software design for performing these functions is described.

The eventual aim of the software design is to produce a graph, in the form of a histogram, of the number of times any real-zero interval occurs. Thus two measurements have to be performed; the interval duration and a count of the number of times it occurs. This is achieved by assigning a memory location to each of the possible interval durations, and incrementing it each time the corresponding duration occurs.

The timing of the interval duration can be performed by detecting the occurrence of the zeros. The simplest way, conceptually, of achieving this is to use the interrupt facility to start and stop the timer. However, an examination of the execution times and the available processing time (50μs) shows that this method is too time consuming and therefore unsuitable.

An alternative technique is to use a "flag" to indicate the occurrence of a zero, and to instruct the microprocessor to perform a periodic inspection of its condition. The setting of the flag can then be used to start and stop the timer, as required. Figure A.1 illustrates the flag inspection procedure. The microprocessor makes the inspections at the times t_1, t_2, t_3, etc and the flag is set at T'. Since this is between the inspections at t_1 and t_2, the earliest time at which the flag can be registered by the microprocessor is during the inspection at t_2. Thus a delay (which is an inaccuracy) of $(t_2 - T')$ seconds results in the detection of the zero-crossing. However, since the resolution of the interval duration is limited to 50μs, the inaccuracy does not present any problems, unless the inspection period exceeds this limit. For measurements on bandlimited speech, therefore, this method is adequate and will be used for subsequent programming.

In addition to the timing and counting functions, two other processes must be included in the complete program - starting and stopping. The starting function includes the initial setting-up of the counters and the timer before beginning the timing and counting operations; and the stopping function includes the output of results.

A.1 The Program

The complete flow diagram of the functions is shown in Figure A.2, where overflow of the counters is used to terminate the program and output the results. In addition, the overflow of the timer is used to indicate the presence of epochs in excess of 255 inspection periods.

The timer is started on detection of the first zero-crossing and incremented on each successive inspection at which the flag is not set (loop 1). On detection of the next zero-crossing, the memory location corresponding to the duration indicated by the timer is incremented and the timer reset (loop 2). Since the execution time of the timing and increment instructions is small, a delay is incorporated to ensure a regular flag inspection at 50μs intervals. Moreover, because the flag inspection can be made from different points in the program, two different delays are provided (loop 1 and loop 2), hence the apparent duplication of part of the program.

The overall structure of this program appears rather complicated and an attempt is made to simplify it by dividing it into routines, as shown in Figure A.3.

A.2 The Routines

Consideration can now be given to the design of each of the routines.

A.2.2 The initialise routine

The purpose of this routine is to set-up the microprocessor for the functions it is to perform. These functions are:

(a) To accept typed-in information and store it in an allocated memory location. This information is the delay data, which is used in determining the delay to be introduced in the timing routine. The data is entered and stored in hexadecimal notation.

(b) To initialise the PIA (Chapter 3 of the applications manual (56)). This entails the setting up of the PIA data and control registers. In this case, the data registers are not used and a simple clearing operation ensures order. In addition, only one of the two control registers is used. The initialisation is thus the setting of the bit pattern on the control register. Bits b_6 and b_7, controlled by inputs CA2 and CA1 respectively, indicate the occurrence of a zero. Bits b_0 - b_5 are pre-set to enable bits b_6 and b_7 to act as the flag bits when signals are connected to CA1 and CA2. The pre-setting is:

b_0 = 0 to disable IRQ1

b_1 = 1 to set b_7 by low-high transition on CA1

b_2 = 1 to select output registers

b_3 = 0 to disable IRQ2

b_4 = 0 to set b_6 by high-low transition on CA2

b_5 = 0 to establish CA2 as input

b_6 = 0)
) to clear flags
b_7 = 0)

Thus, loading CRA with $06 achieves the necessary initialisation.

(c) To clear the memory locations. It is necessary to use 2 x 256 memory locations as counters. To ensure an initially zero condition on all counters, a clearing operation is performed on the relevant memory locations.

(d) To detect the first zero (PIA flag). The inputs CA1 and CA2 control the status of bits b_6 and b_7. A positive transition on CA1 sets b_7 and a negative transition on CA2 sets b_6. Since it is only the status of bits b_6 and b_7 which indicate the occurrence of a zero, a BIT operation is performed to see if a zero has occurred. BITB instruction performs an AND operation on ACCB and memory. If accumulator B is loaded with the contents of the control register, CRA, and the AND operation performed with $CO, the presence of a flag is indicated by a non-zero result. For example,

LDA B	PIACRA	b_7	b_6	X	X	X	X	X	X
	$CO	1	1	0	0	0	0	0	0
BIT B	#CO	b_7	b_6	0	0	0	0	0	0

When a flag is detected it is reset before proceeding further, so that the occurrence of successive flags can be detected. This is accomplished by a dummy read operation on the output register, as:

```
LDA  B    PIAORA
```

The flow diagram of this routine is shown in Figure A.4.

A.2.2 The timing routine

The purpose of this routine is to time the duration between successive zero-crossings and to increment the appropriate memory locations. The steps are:

(a) To increment the time measuring device, which in this case is Accumulator A.

(b) To check the status of the carry bit after incrementing accumulator A. This determines whether or not the count on accumulator A has exceeded 255. If the carry bit is set the microprocessor is instructed to exit the timing routine and enter "the record 256" subroutine.

(c) To cause a delay. The purpose of the delay is to allow sufficient time to elapse before the next inspection of the PIA flag.

(d) To inspect the PIA flag.

(e) To increment the memory which corresponds with the interval duration indicated by the contents of accumulator A. This operation is only performed if the PIA flag has been detected within the routine.

Since the contents of accumulator A can at most be 255, the longest interval duration to be measured is the time for 255 periods of the inspection frequency. Thus the maximum number of different interval durations is 255. The simplest way of determining the counter which corresponds to the contents of accumulator A is to use memory locations 1-255 as the counters. In this way the memory to be incremented is given directly by the count on accumulator A. That is, if the contents of accumulator A when the flag is detected is ten, then memory location ten is incremented.

(f) To check the status of the carry bit after incrementing a memory location. This is to check the counter for overflow, upon which the microprocessor is instructed to leave the timing routine and enter "the secran" subroutine.

Figure A.5 shows the flow diagram for the timing routine.

A.2.3 The secran subroutine

This subroutine extends the counters to sixteen bit working. Eight bit counters allow about one second of

speech to be played before one of them becomes overloaded and terminates the program. Extension to sixteen bit working allows about ten seconds at the very least, but a speech sample of some minutes can be played before overloading any of the counters. In practice it is found that speech lasting four or five minutes can be easily accommodated.

The functions of this routine are:

(a) To increment a memory location in the second range when the corresponding memory in the first range becomes overloaded.

(b) To jump to "the exit" subroutine when any of the memory locations in the second range overflow.

(c) To introduce a delay before returning to the timing routine.

Figure A.6 shows the flow diagram.

A.2.4 The record 256 subroutine

This subroutine indicates, by incrementing memory 256, that the interval duration has exceeded 255 periods of the inspection frequency. It is not very easy to measure interval durations longer than 255, nor is it a very useful measurement. In such cases, therefore, it is sufficient to say that an interval duration has exceeded

this limit. The flow diagram for this subroutine is shown in Fig A.7, and it is not necessary to give a detailed description of its functions since they are self-explanatory.

A.2.5 The exit subroutine

This subroutine is the most intricate in design, but its basic function is to present the contents of the counters in a form which is recognisable by the X-Y plotter. This is achieved by performing the following steps:

(a) To determine whether or not the second range of memories has been used; and if so, to perform a sixteen to eight bit conversion. When using both the memory ranges, the counters are of sixteen bit length, but the data bus is only capable of handling eight bit words. Thus, before transmitting the information about the contents of the counters via the data bus, a sixteen to eight bit conversion is necessary. This conversion is achieved by first determining the counter with the highest contents and then rotating the sixteen bits until the MSB is in position b_{15}. A record is kept of the number of positions the MSB is rotated before it reaches the position b_{15}. All other counters are rotated the same number of positions. Finally, the eight most significant bits of the sixteen are used to represent the count.

(b) The conversion of hexadecimal to decimal notation. The microprocessor uses hexadecimal notation, but it is more convenient in the outside world to use decimal notation. Therefore, a conversion is necessary when transmitting to the outside world.

(c) The translation to the X-Y plotter format. Although the data is converted to decimal notation, it is still not in a suitable form for the X-Y plotter to recognise. Therefore it is necessary to translate the contents of the counter and the information about the counter identification into X-Y plotter recognisable format. When the translation is complete it is transmitted over the data bus to the X-Y plotter memory box. The X-Y plotter control then takes over and causes the plotter to print the contents of the counter against the counter position.

The flow diagram for this subroutine is shown in Figure A.8.

A.3 Testing the Software

Having designed the program it is necessary to test its operation to ensure that it gives the correct response. This is achieved by performing the probability measurements on a signal of known parameters.

A sine wave of a known frequency is expected to give a peak at the interval duration corresponding to the known

frequency. Performing the tests on a sine wave with a half period of 0.5ms in fact resulted in a graph which showed a peak at 0.5ms and a smaller peak at the interval adjacent to 0.5ms.

After detailed consideration of the expected results, it was concluded that the results obtained are in fact correct. An explanation to this effect now follows.

Figure A.9 shows the square wave input to the microprocessor, of period 2 x Tp secs. The duration of the timing cycle in the microprocessor is Ti seconds and the number of timing cycles it takes to detect the zero-crossings depends on the length of Tp. If Tp is an integer multiple of Ti, then the number of timing cycles necessary to detect the zero-crossings would be equal to the value of the integer. If, however, Tp is such that:

$$NTi < Tp < (N + 1)\,Ti$$

then the number of cycles to detection is not N or (N + 1) exclusively, but a combination of both. This can be illustrated by an example. Let $Tp = (N + \Delta)Ti$ and let $\Delta = 0.25$. It would take (N + 1) cycles to detect the first zero. That is, 0.75 cycles of the next period have already elapsed. The interval to the next zero is $(N + \Delta) - 0.75 = N - 0.5$. Therefore, it would take only N cycles to detect the second zero. Construction of the chart (table A.1)

below shows how the number of intervals to detection interchanges between N and N + 1. Continuation of the chart shows a pattern emerging. The sequence repeats itself after every four periods; one out of every four periods takes N + 1 cycles to detect. The ratio,

$$\frac{\text{Number of periods it takes N+1 cycles to detect}}{\text{Total number of periods}} = \frac{1}{4} = \Delta$$

Therefore, Δ determines the ratio of the two adjacent peaks in the distribution graph. This is verified practically by a sine wave with a half period of 210μsecs, which gives the results shown in Figure A.10; peaks at 200μs and 250μs. In this case Δ is 0.2 and the peak at 250μs is approximately a fifth of the sum of the two peaks.

Period Number	Number of intervals to detection	Excess	Interval to next zero
1	N + 1	0.75	(N+Δ) - 0.75 = N-0.5
2	N	0.50	(N+Δ) - 0.5 = N-0.25
3	N	0.25	(N+Δ) - 0.25 = N
4	N	0	(N+Δ) = N+0.25
5	N + 1	0.75	(N+Δ) - 0.75 = N-0.5
6	N	0.50	(N+Δ) - 0.5 = N-0.25
7	N	0.25	(N+Δ) - 0.25 = N
8	N	0	(N+Δ) = N+0.25
9	N + 1	0.75	(N+Δ) - 0.75 = N-0.5
'	'	'	'
'	'	'	'
'	'	'	'

TABLE A.1

APPENDIX B

CHARACTERISTICS OF THE DIGITISED SPEECH FILES

Five digitised speech files are available for processing on the PDP-11 computer. Table B.1 lists the general characteristics of these files.

TABLE B.1

Speech file	Utterance	Bandwidth	RMS	Peak	File size	Duration of utterance	Type of speaker	No of epochs
- .SPH		Hertz			Blocks	Msecs	M/F	
APPLE7	An apple a day keeps the doctor away	10-4500	237	1449	75	1920	M	1821
APPLE8	An apple a day keeps the doctor away	10-4500	418	1840	86	2202	M	1925
CBONLY	Charles bottleneck	10-4500	187	1579	76	1946	M	4339
BIRD	A bird in the hand is worth two in the bush	10-4500	450	2045	100	2560	M	2727
FEM1	That's very nice for swimming	10-4500	174	1094	69	1766	F	2079

F I G U R E S

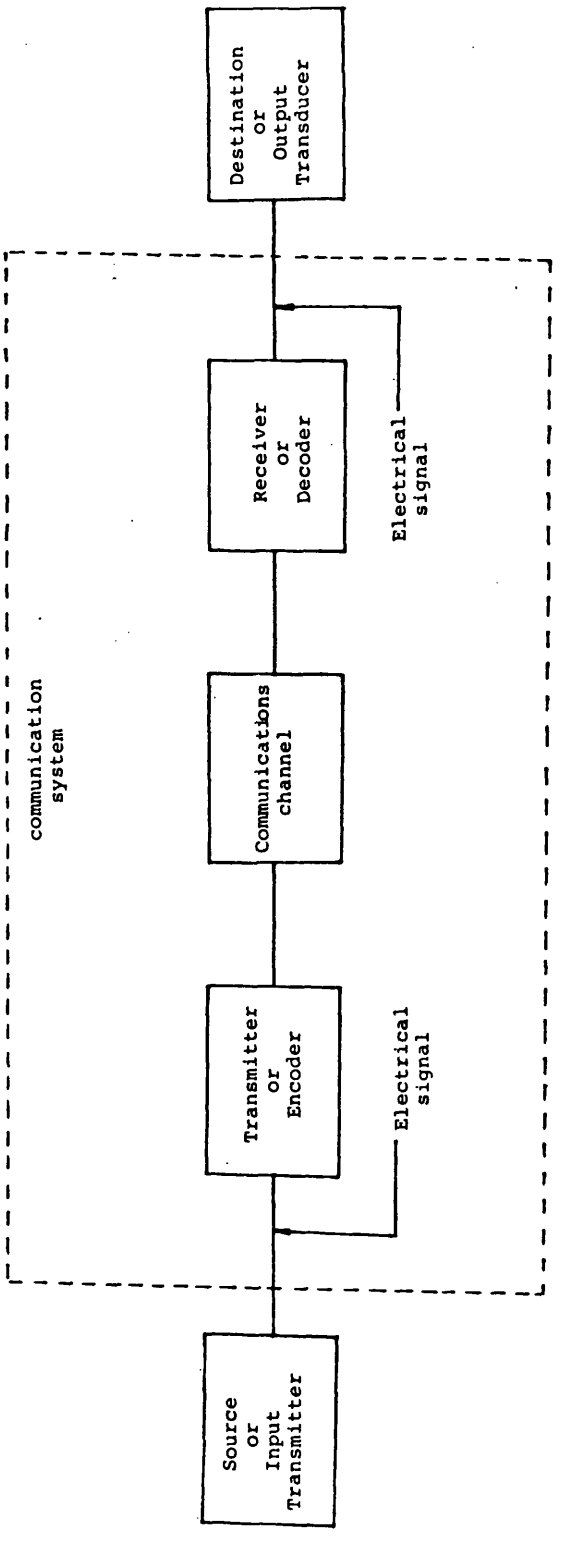


Figure 1.1 The Model of an Electrical Communication System

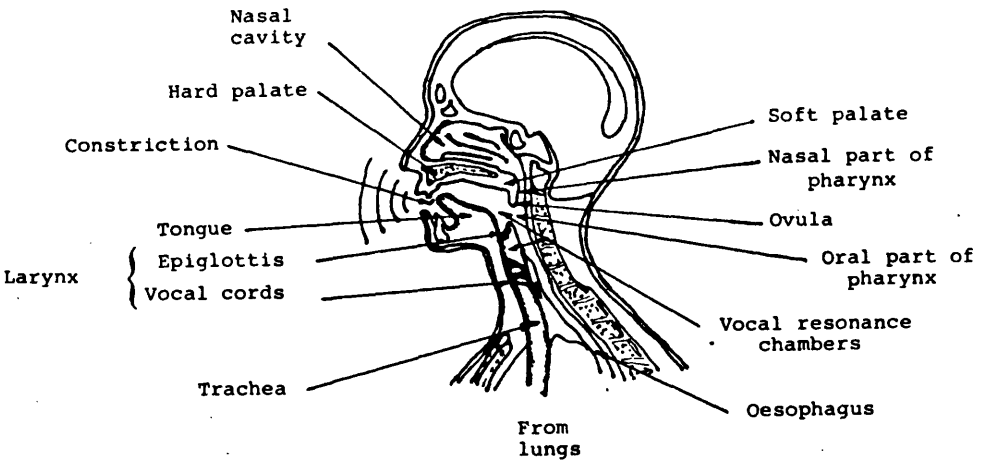


Figure 1.2 Schematic diagram of the human vocal system

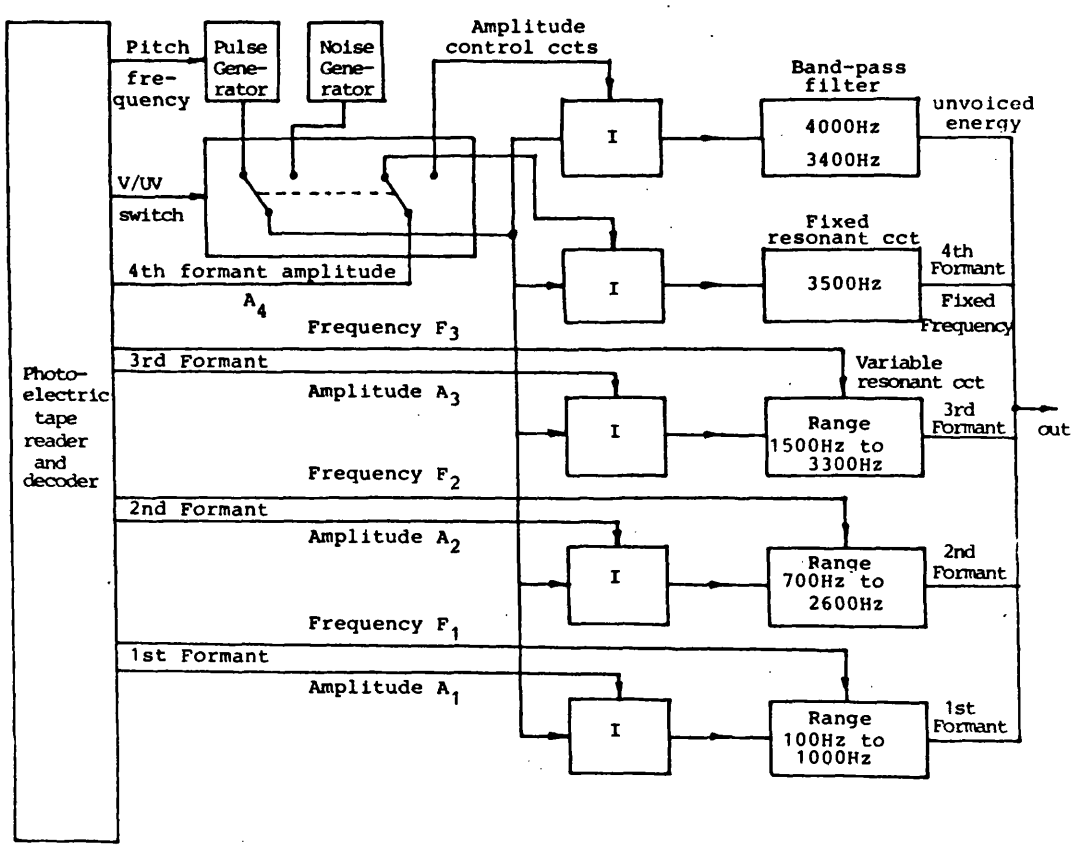


Figure 1.3 Block diagram of a formant synthesizer

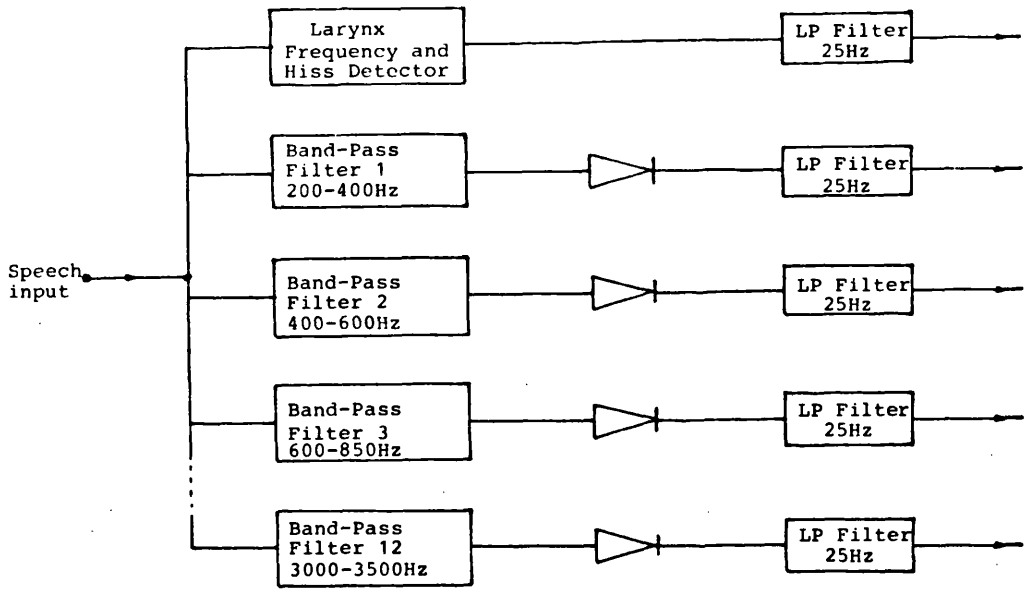


Figure 1.4

Channel Vocoder Analyser

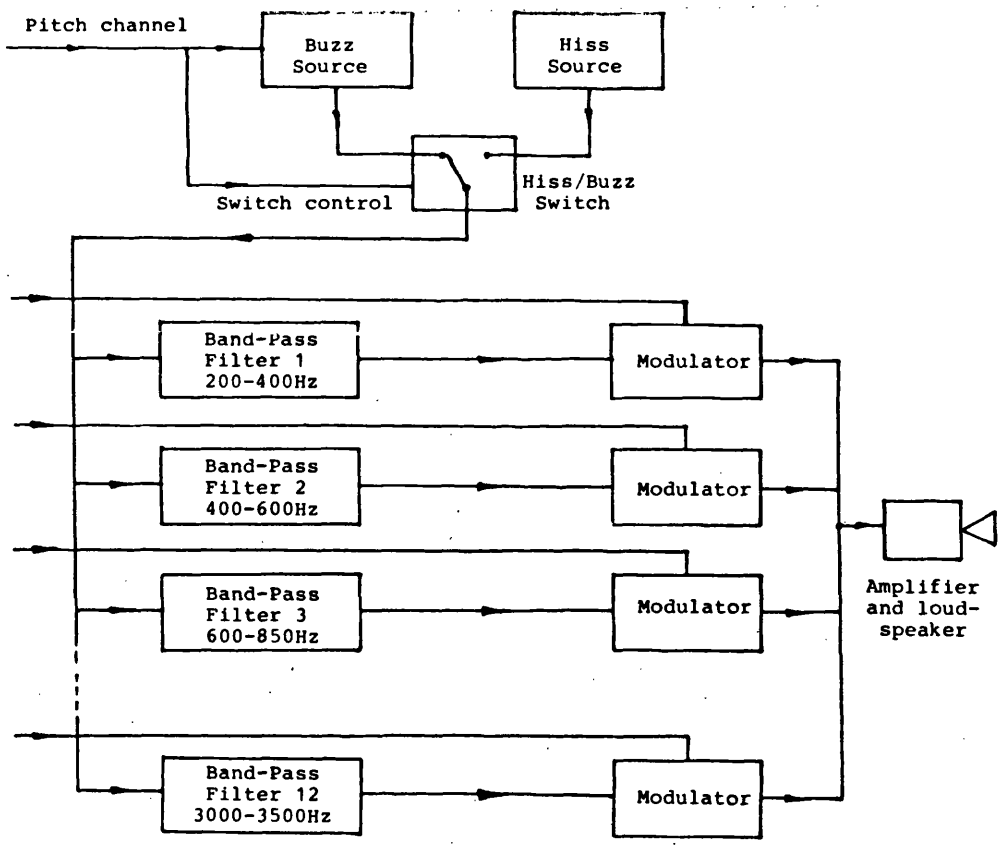


Figure 1.5

Channel Vocoder Synthesiser

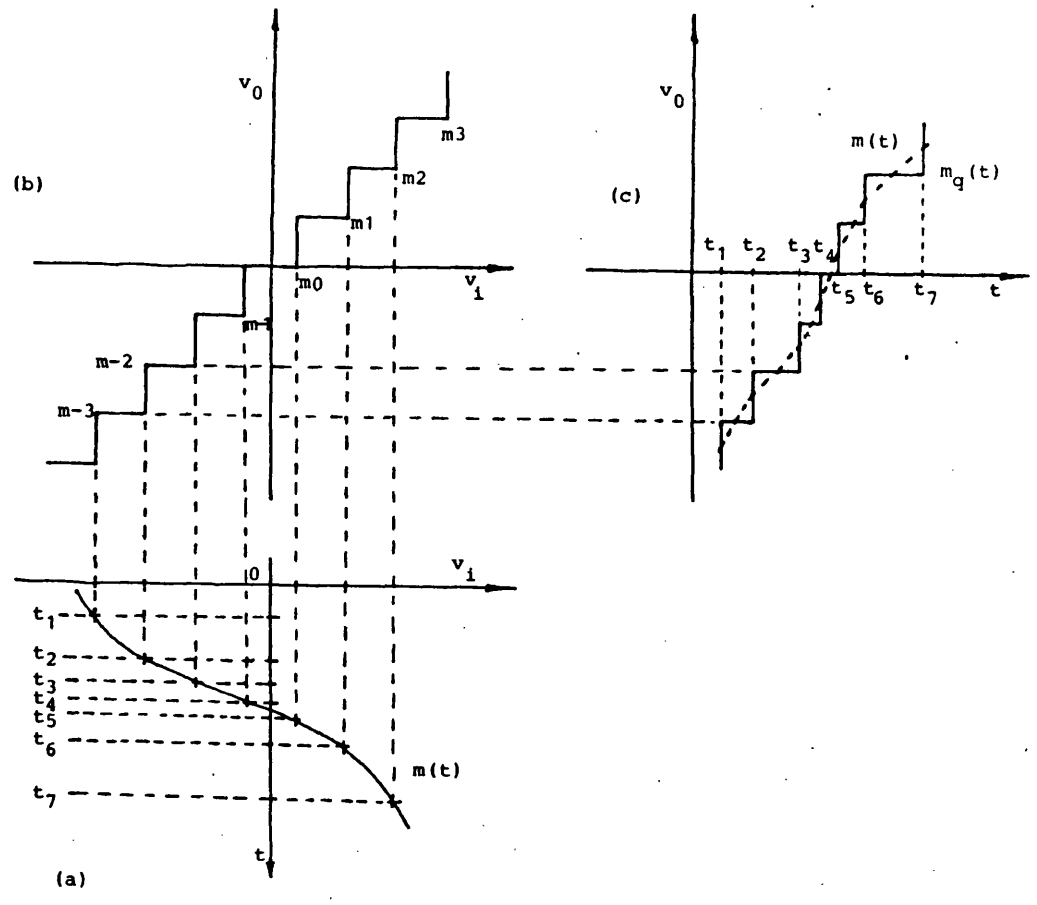


Figure 1.6

Illustrating quantisation: (a) input signal m(t);
(b) Quantiser input-output characteristic;
(c) Quantiser output $m_q(t)$

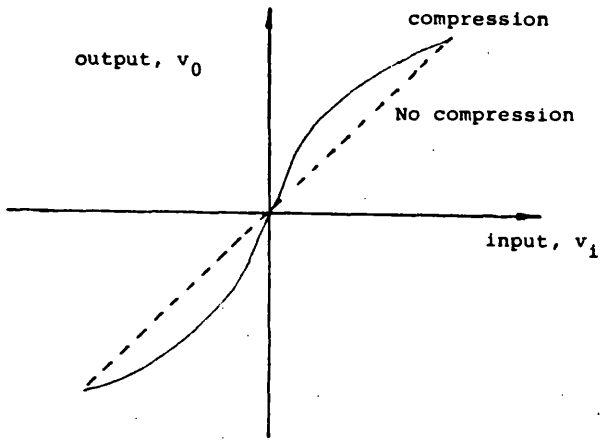


Figure 1.7

The input-output characteristic of a compressor

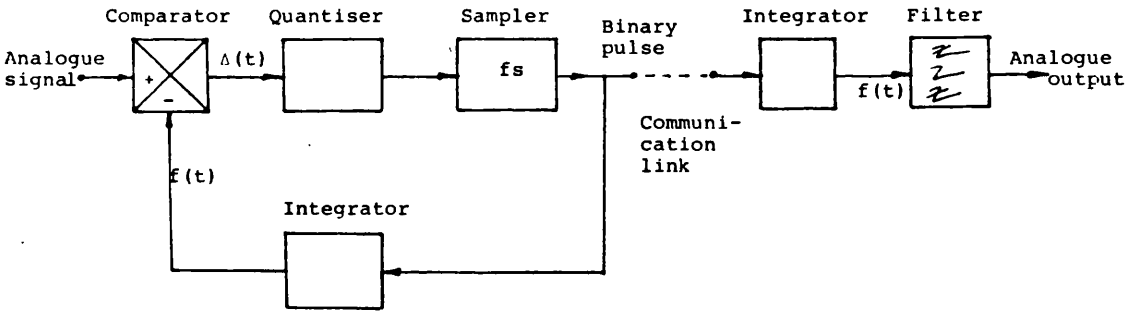


Figure 1.8

Basic Delta Modulator System

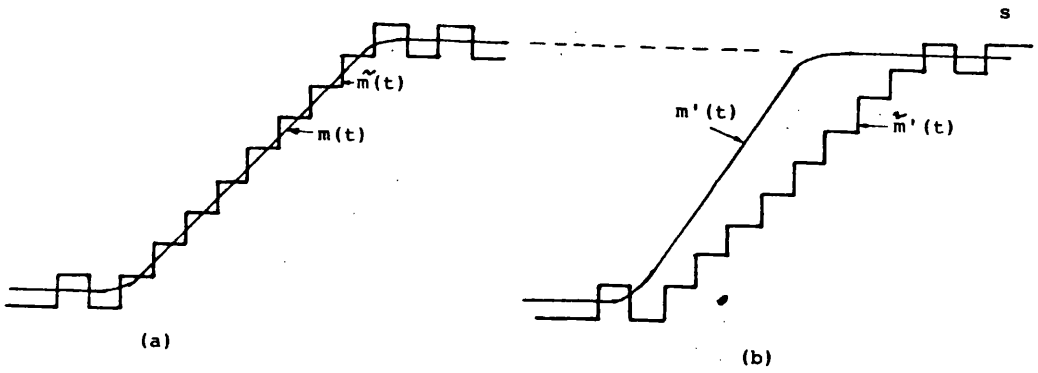


Figure 1.9

Illustration of slope overload: (a) shows $\tilde{m}(t)$ matching $m(t)$; (b) shows $\tilde{m}'(t)$ unable to match the greater rise in $m'(t)$.

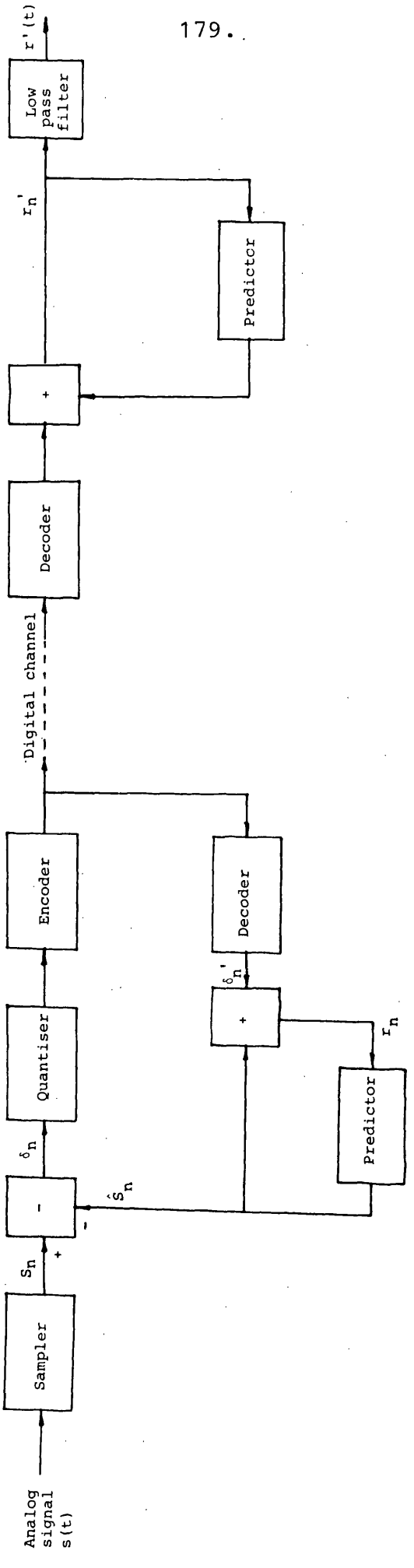


Figure 1.10 Block diagram of a predictive coding system

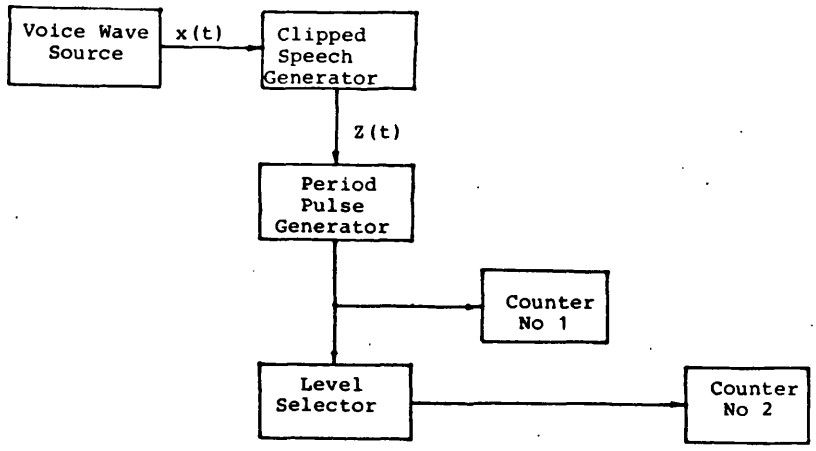


Figure 2.1

Davenport's Zero-crossing distribution measuring equipment

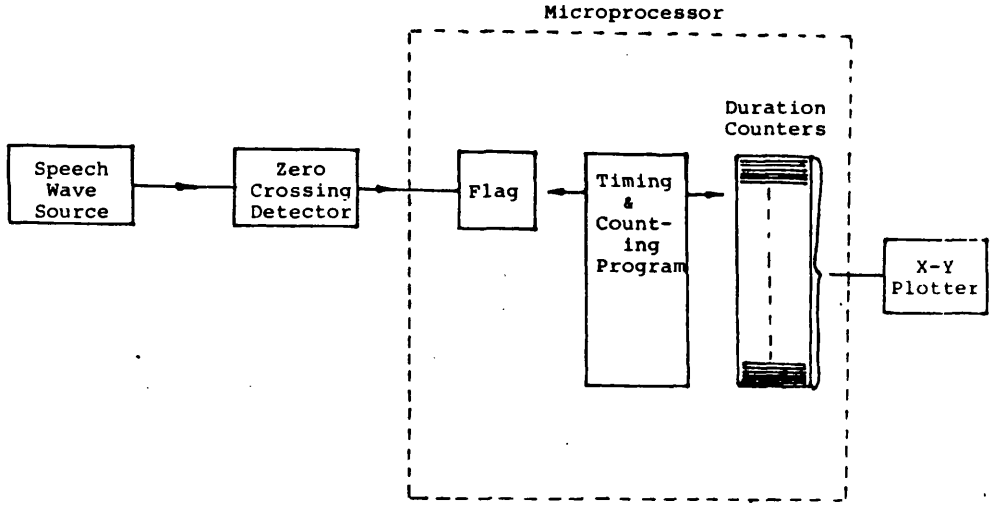


Figure 2.2

The Microprocessor Method Distribution Measuring Equipment

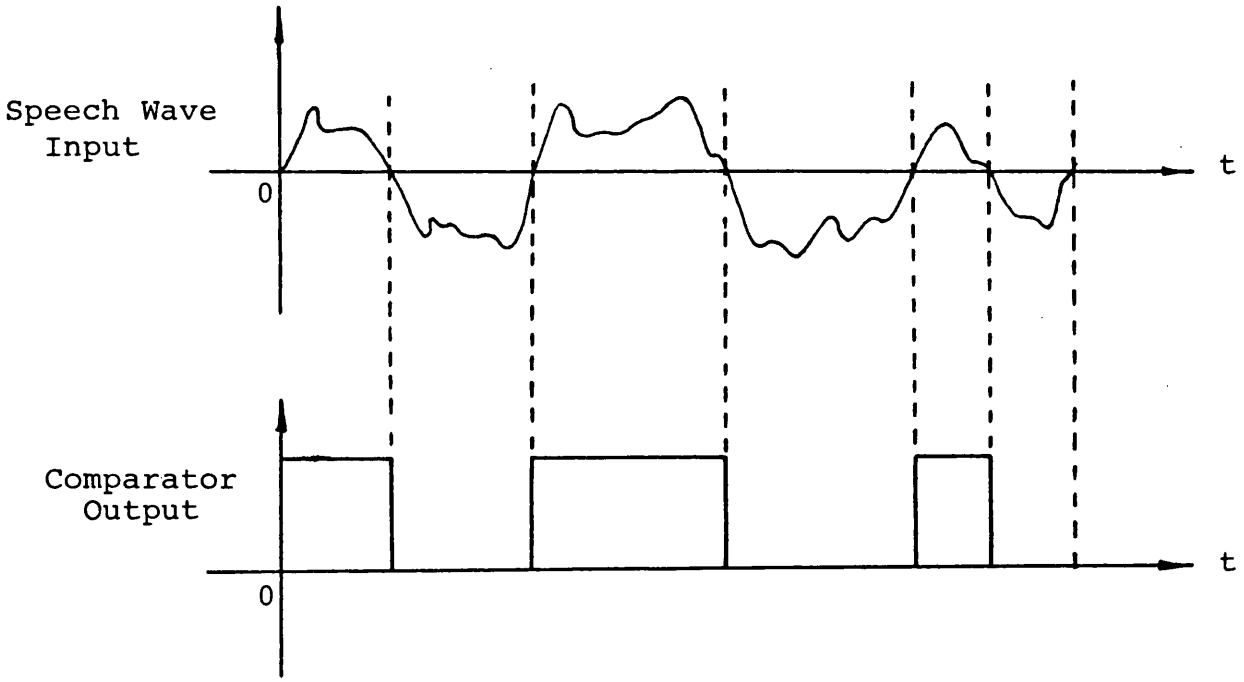


Figure 2.3 Typical zero-crossing detector input and output waveforms

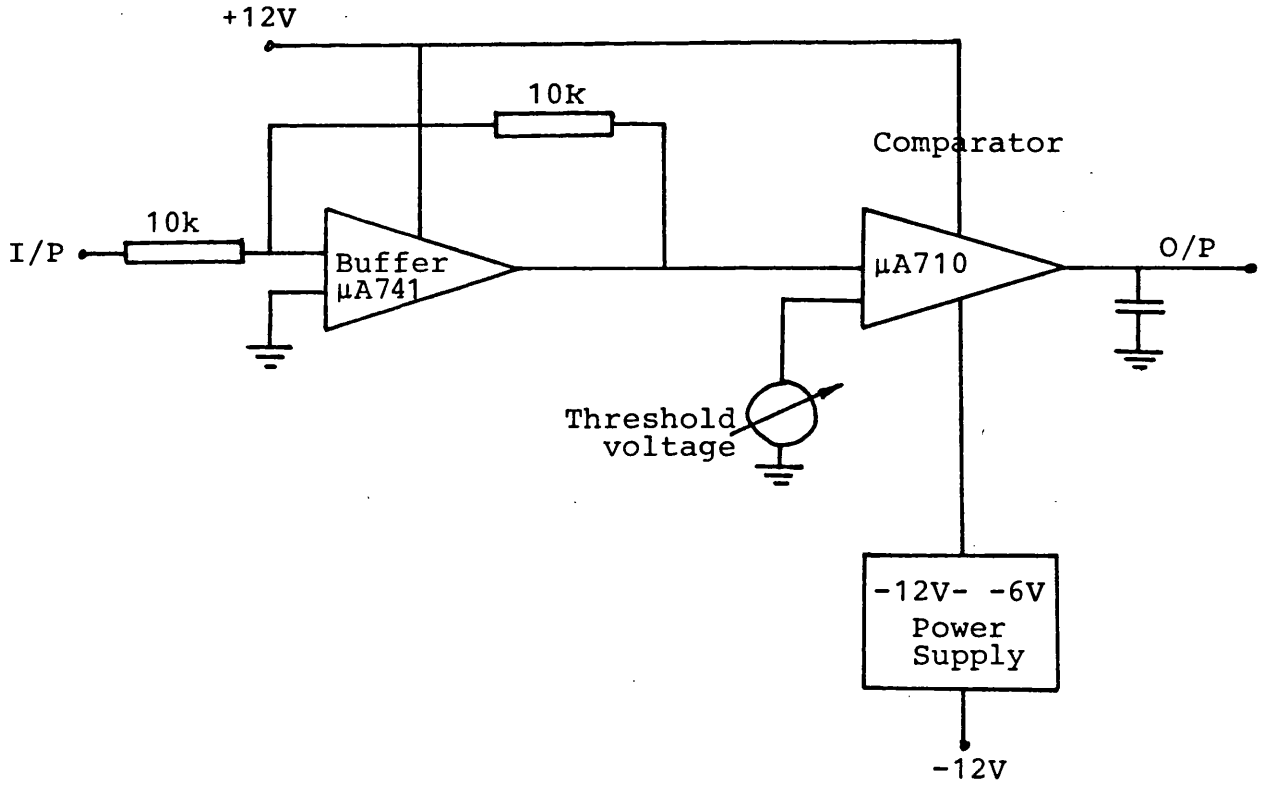


Figure 2.4 The zero-crossing detector CCT

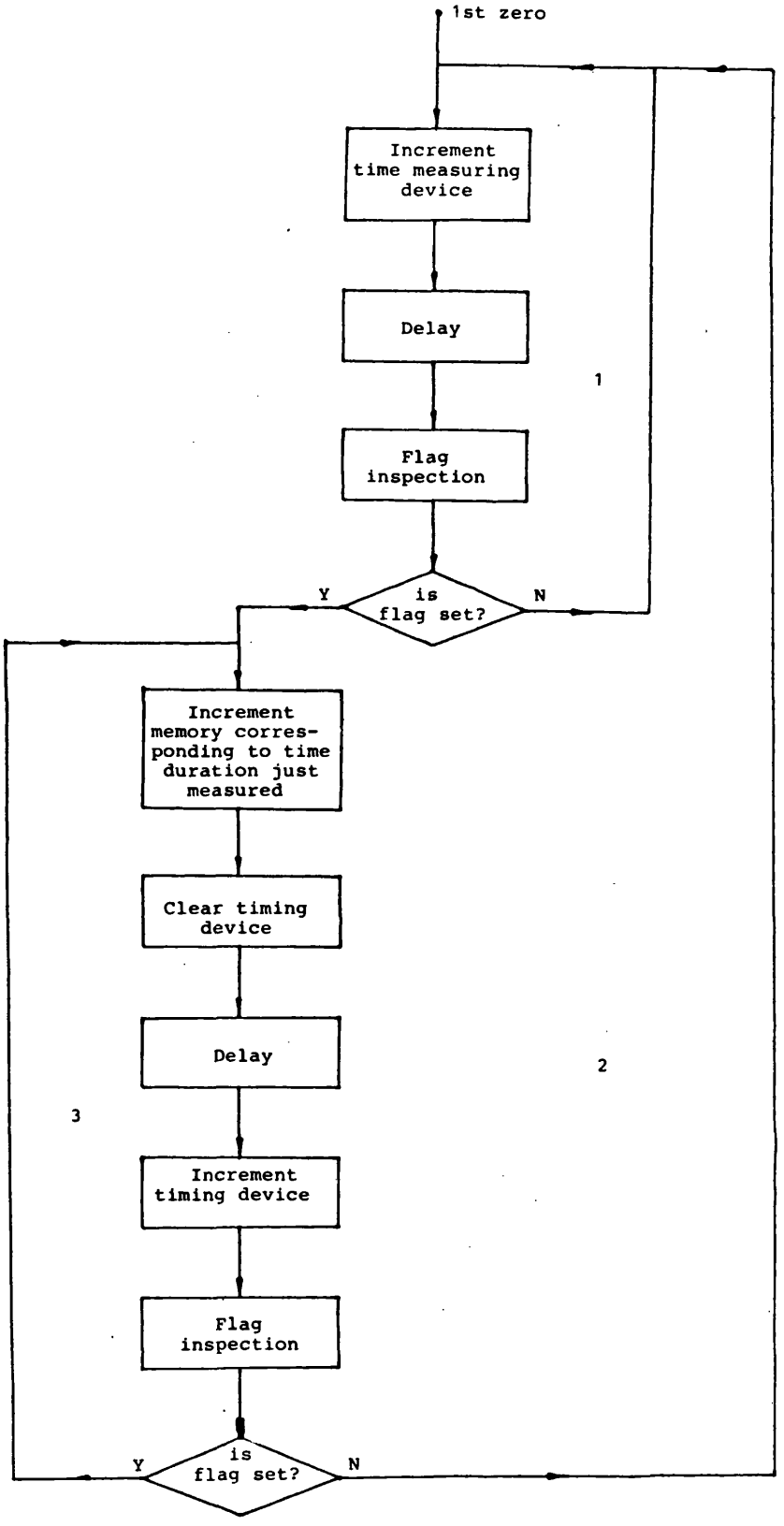


Figure 2.5

The Timing and Counting Flow Diagram

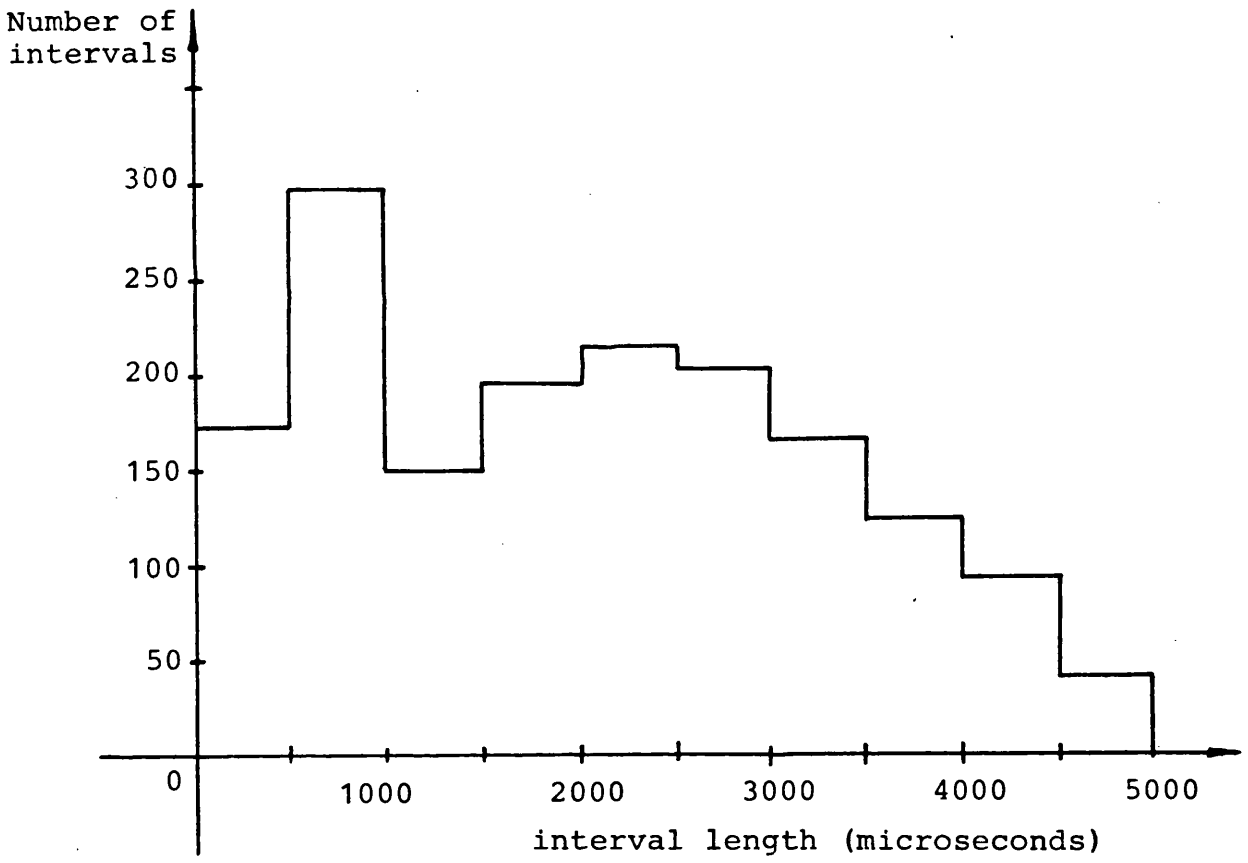
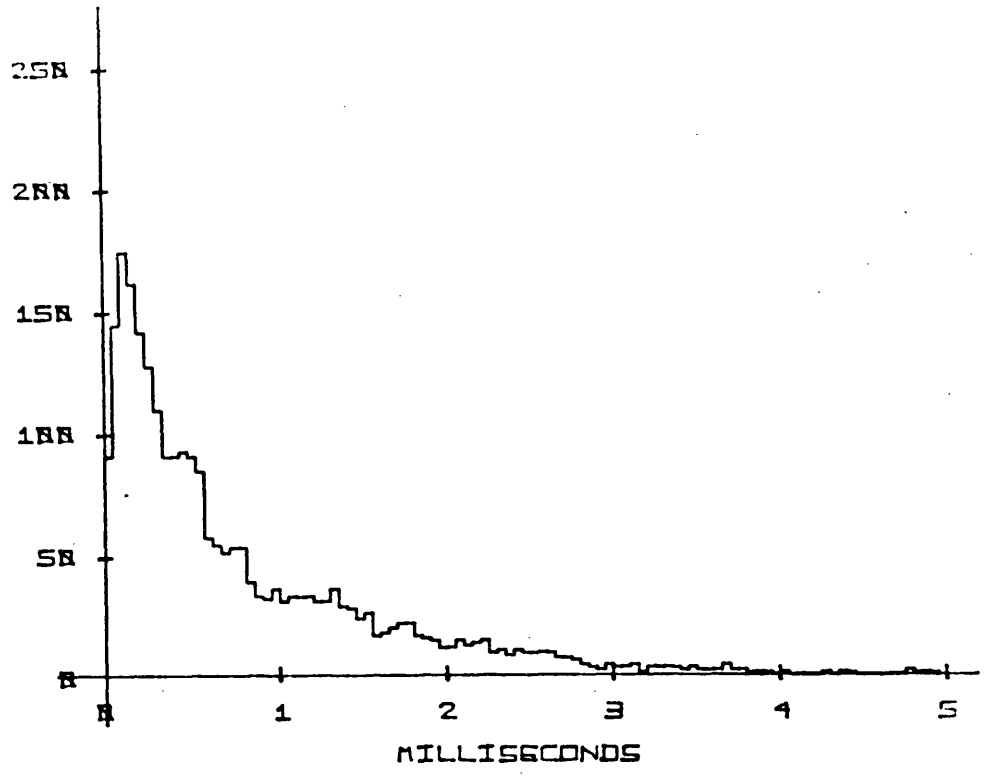


Figure 2.6 Histogram format of the microprocessor results

(a)



(b)

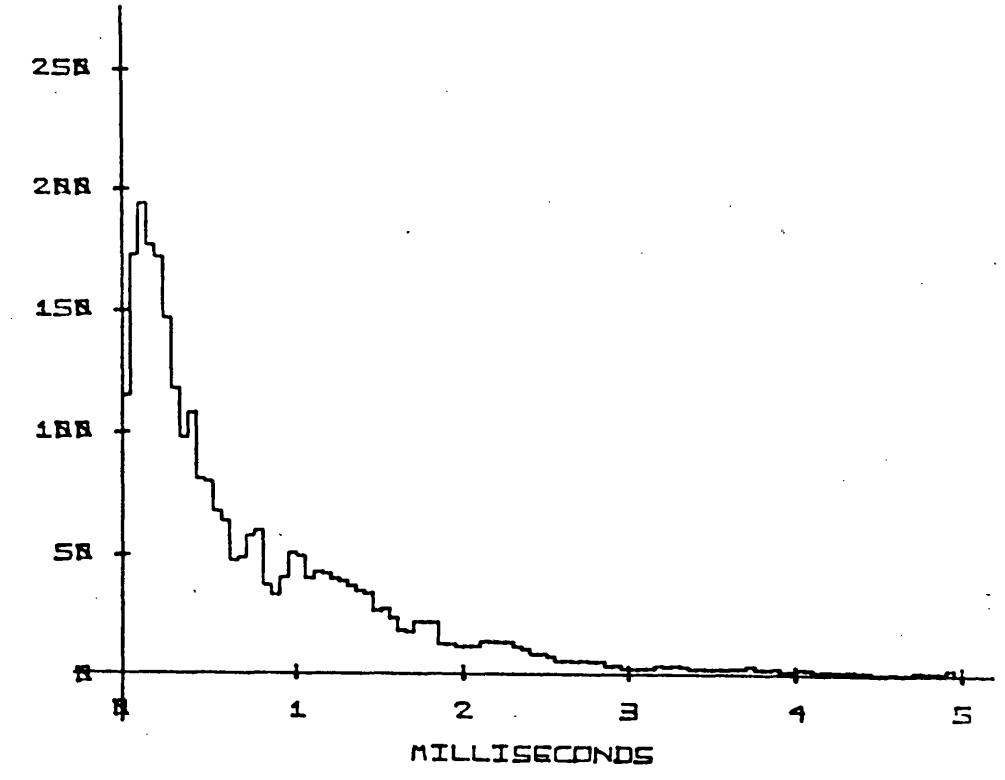


Figure 2.7 The real-zero probability distribution, illustrating the use of Davenport's bias signal and the threshold voltage.

(a) 20Hz square wave bias signal with amplitude 100mV

(b) DC threshold voltage with amplitude 100mV

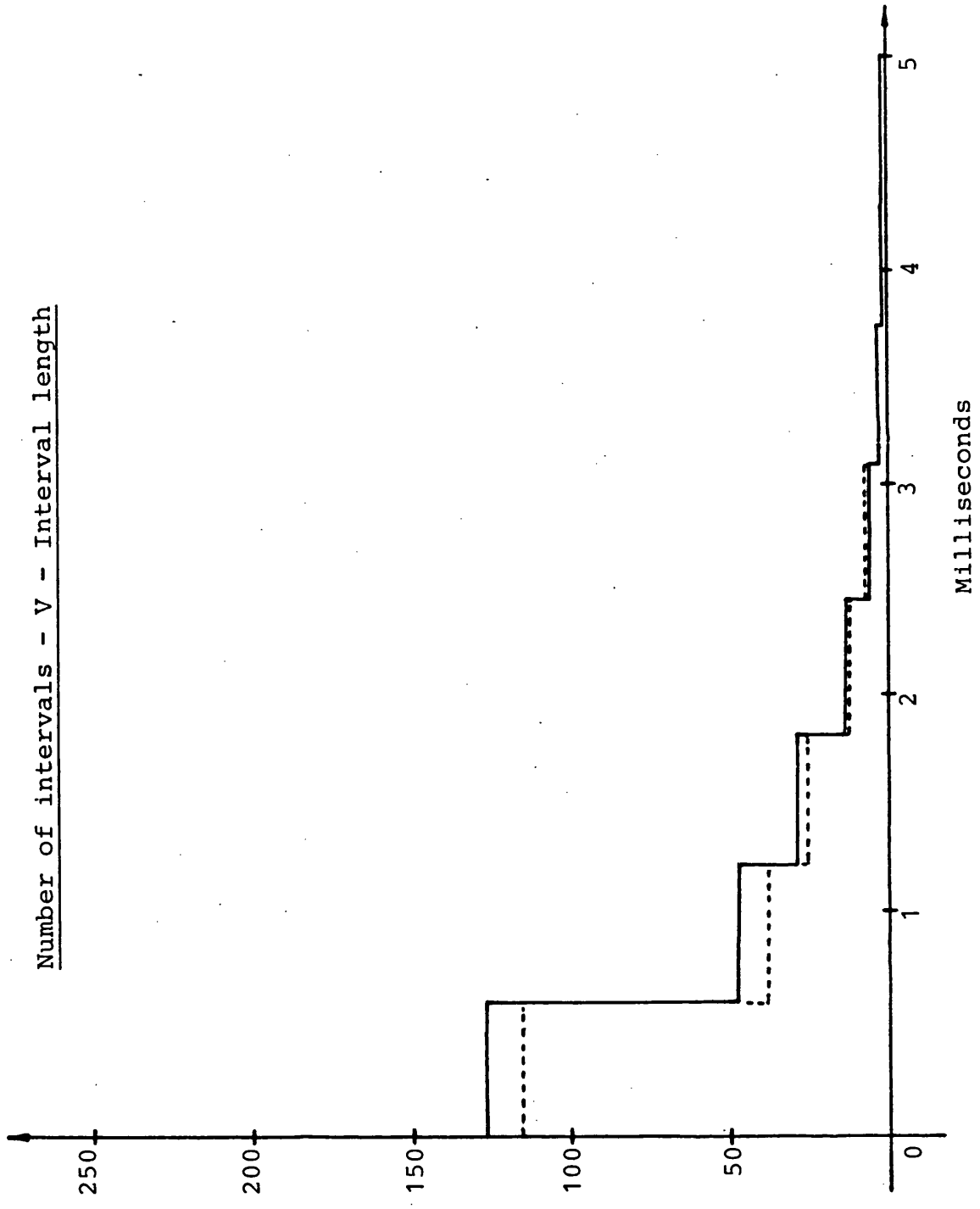
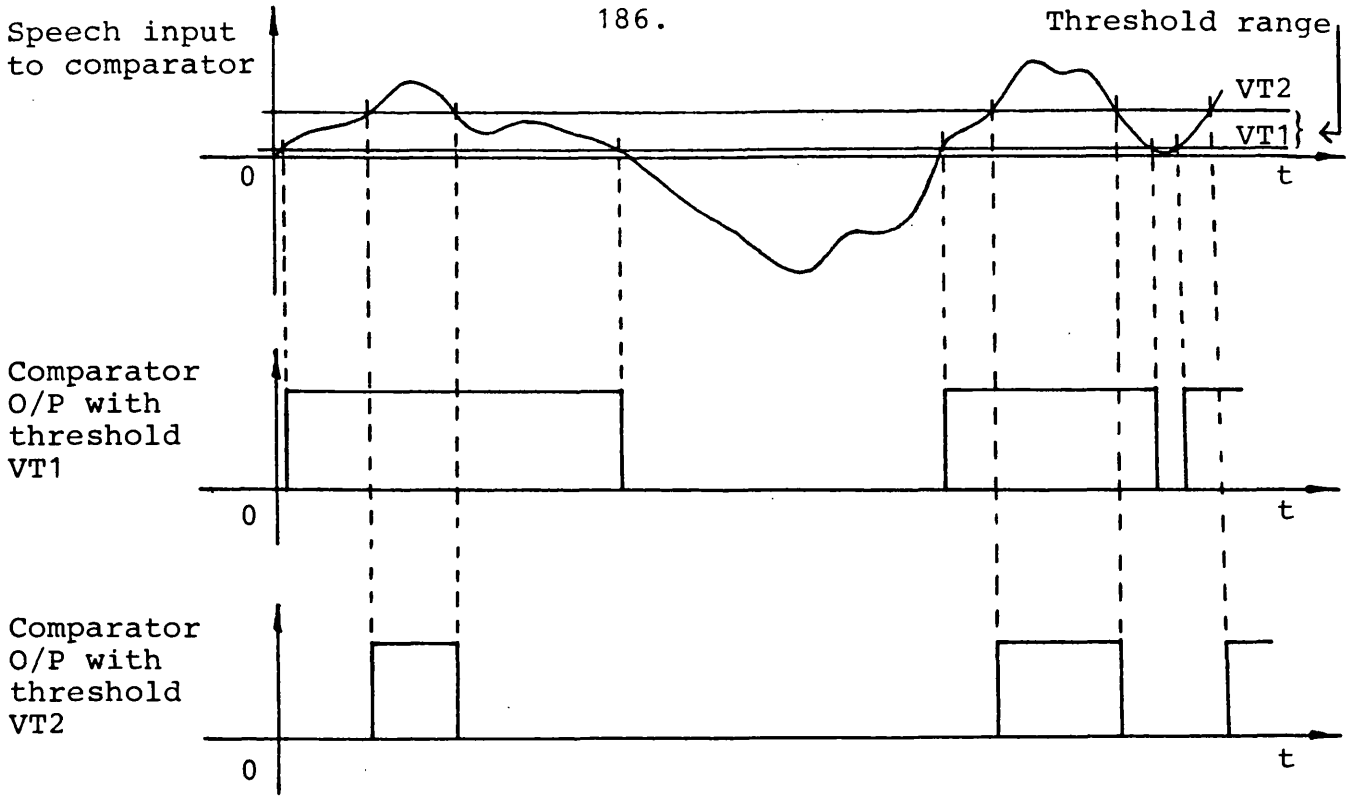


Figure 2.8 The real-zero probability distribution employing Sturge's rule of interval reduction

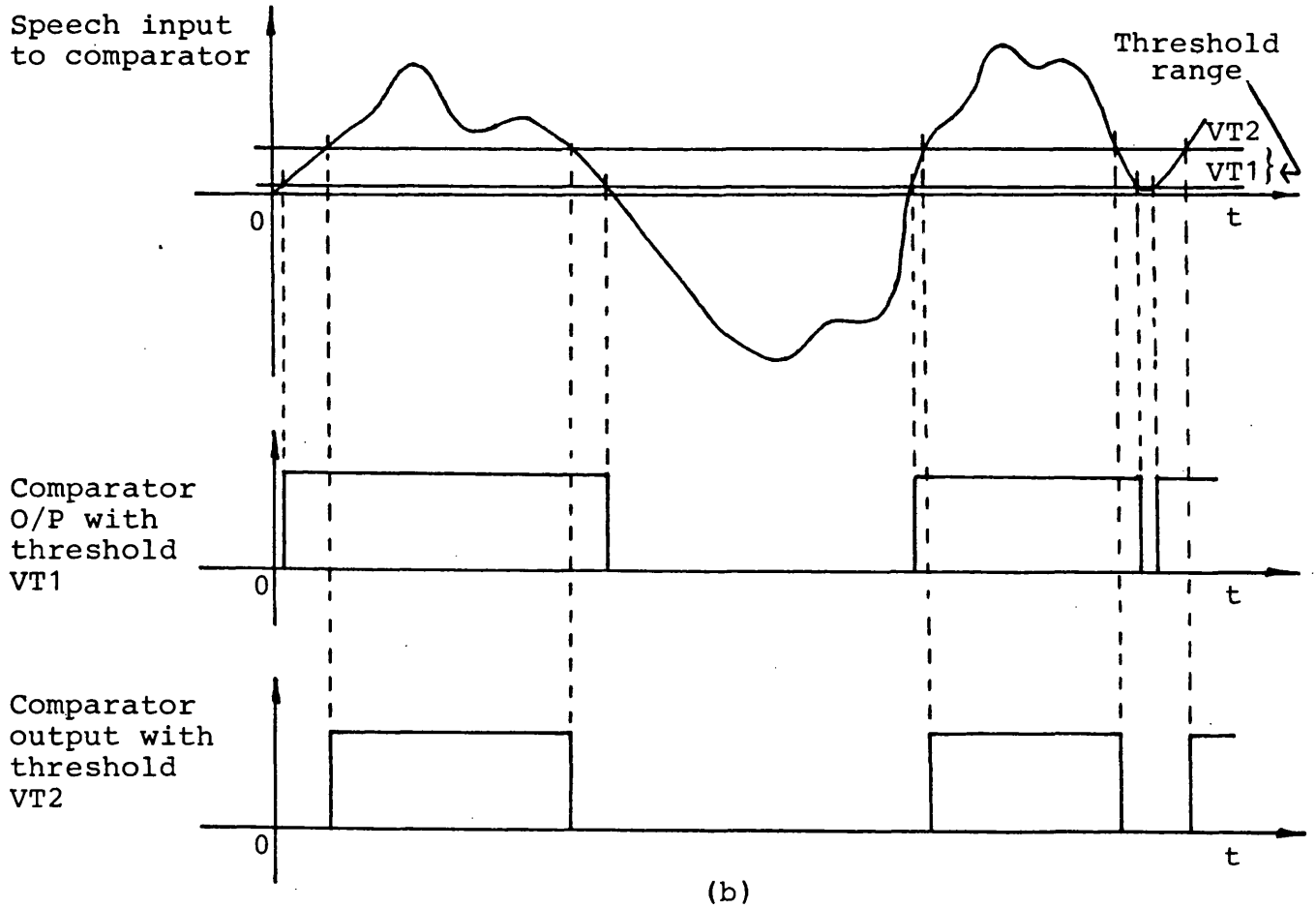
Illustrates similarity of the threshold voltage and Davenport's bias signal.

_____ DC threshold voltage of 100mV amplitude

------ Davenport's 20Hz square wave bias signal of 100mV amplitude



(a)



(b)

Figure 2.9 The effect of low and high amplitude input on comparator output

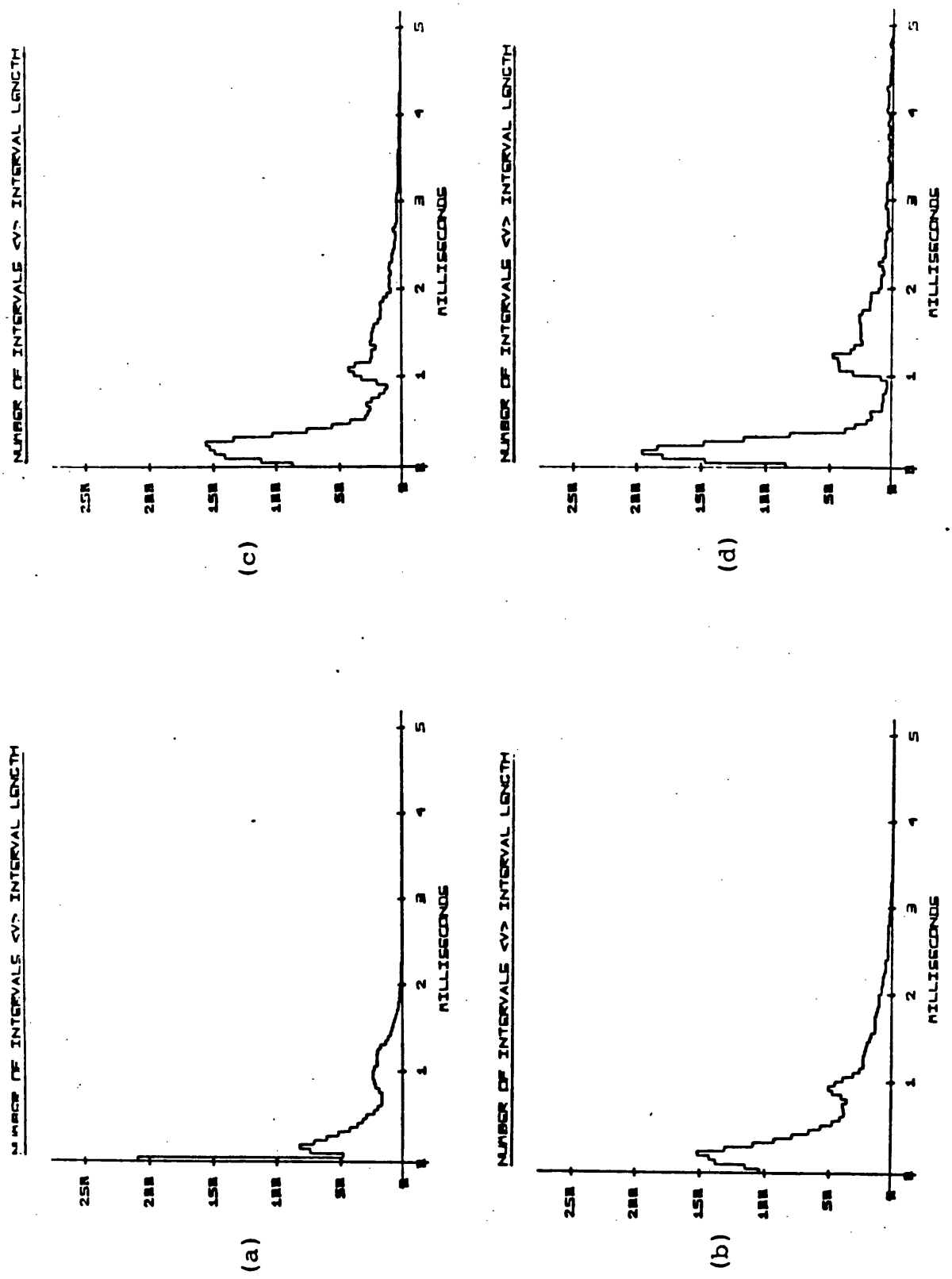
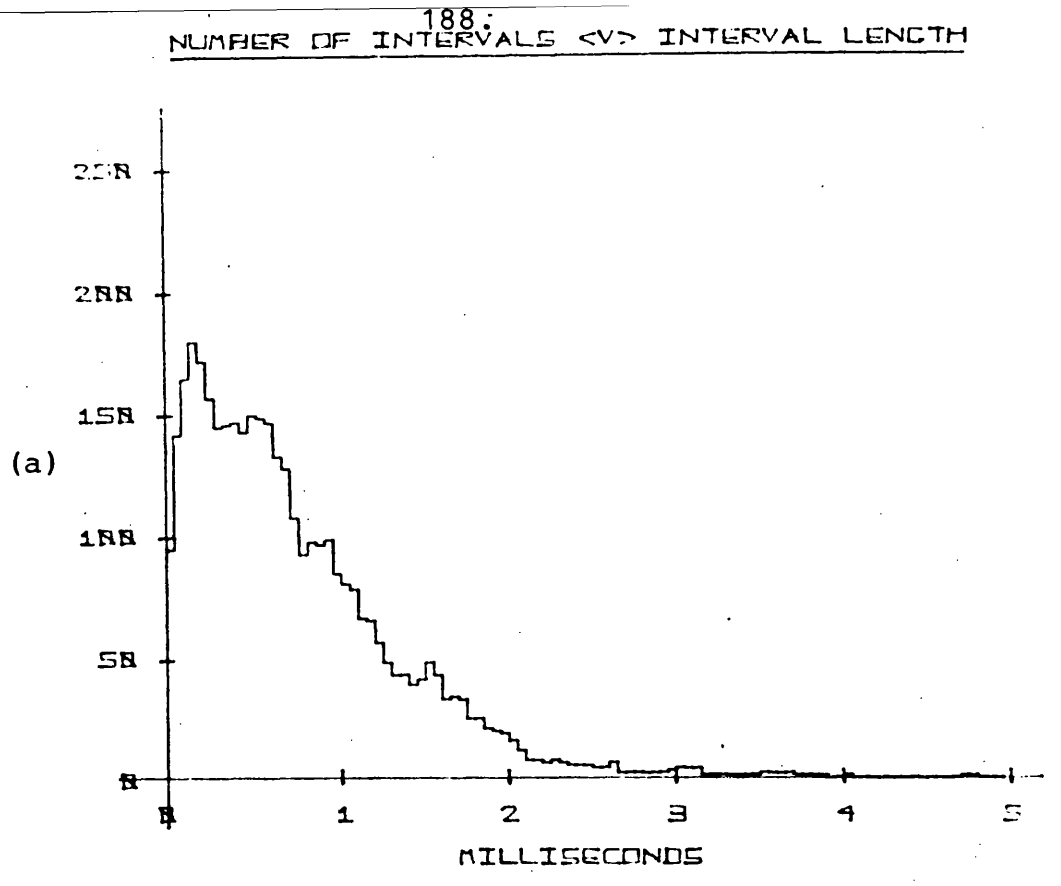


Figure 2.10 The real-zero probability distributions of speech, illustrating the effect of threshold variation.

(a) Threshold : 0mV (b) Threshold : 40mV

(c) Threshold : 80mV (d) Threshold : 120mV

Bandwidth: 100-5000Hz



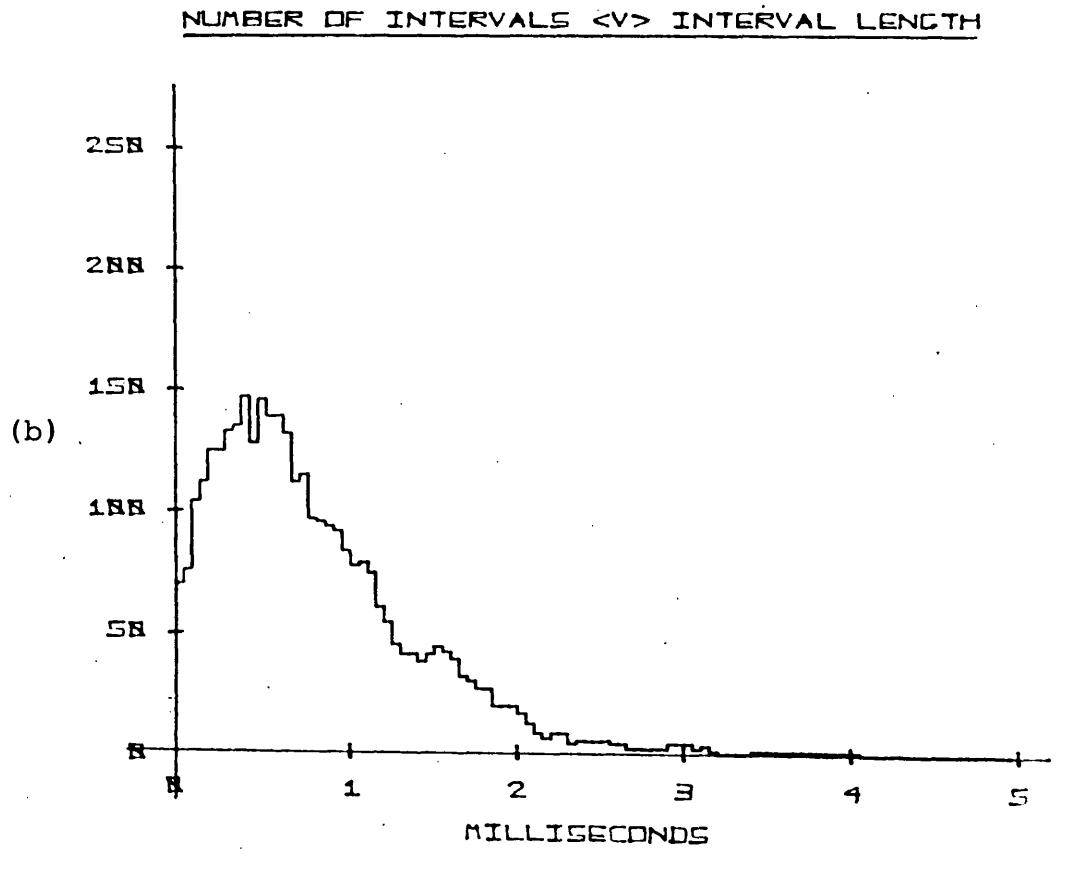
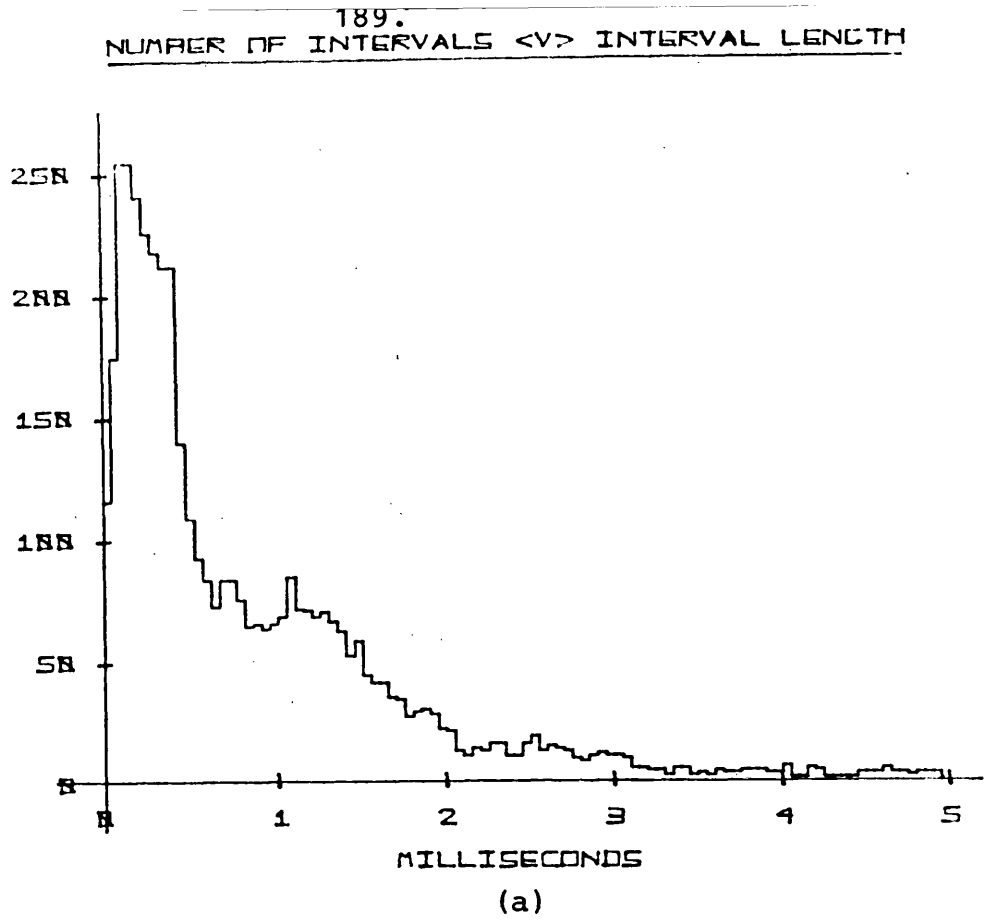


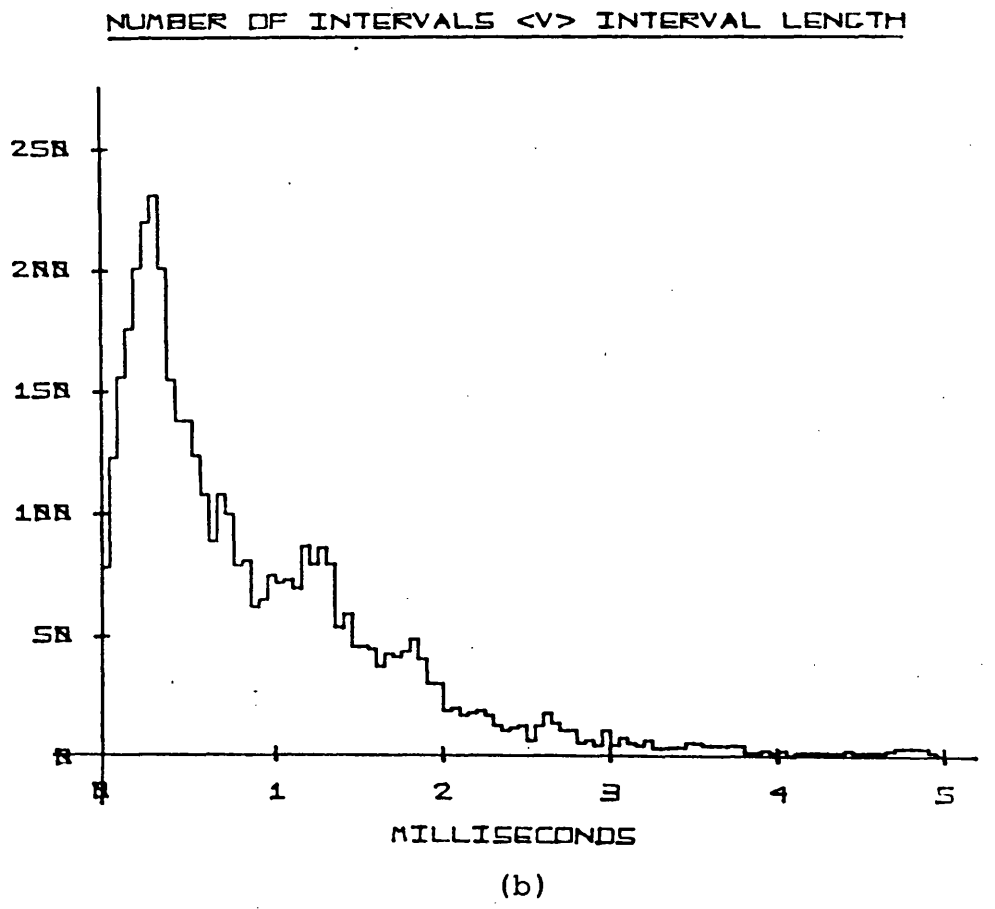
Figure 2.11 The real-zero probability distribution of speech, illustrating the reduction of high frequency components with reduction of the bandwidth (threshold : 60mV)

(a) bandwidth : 300-5000 Hz

(b) bandwidth : 300-3400 Hz



(a)



(b)

Figure 2.12 The real-zero probability distribution of speech, illustrating the effect of bandlimiting

(a) Threshold : 120mV; Bandwidth : 100-5000Hz;

(b) Threshold : 120mV; Bandwidth : 300-3400Hz.

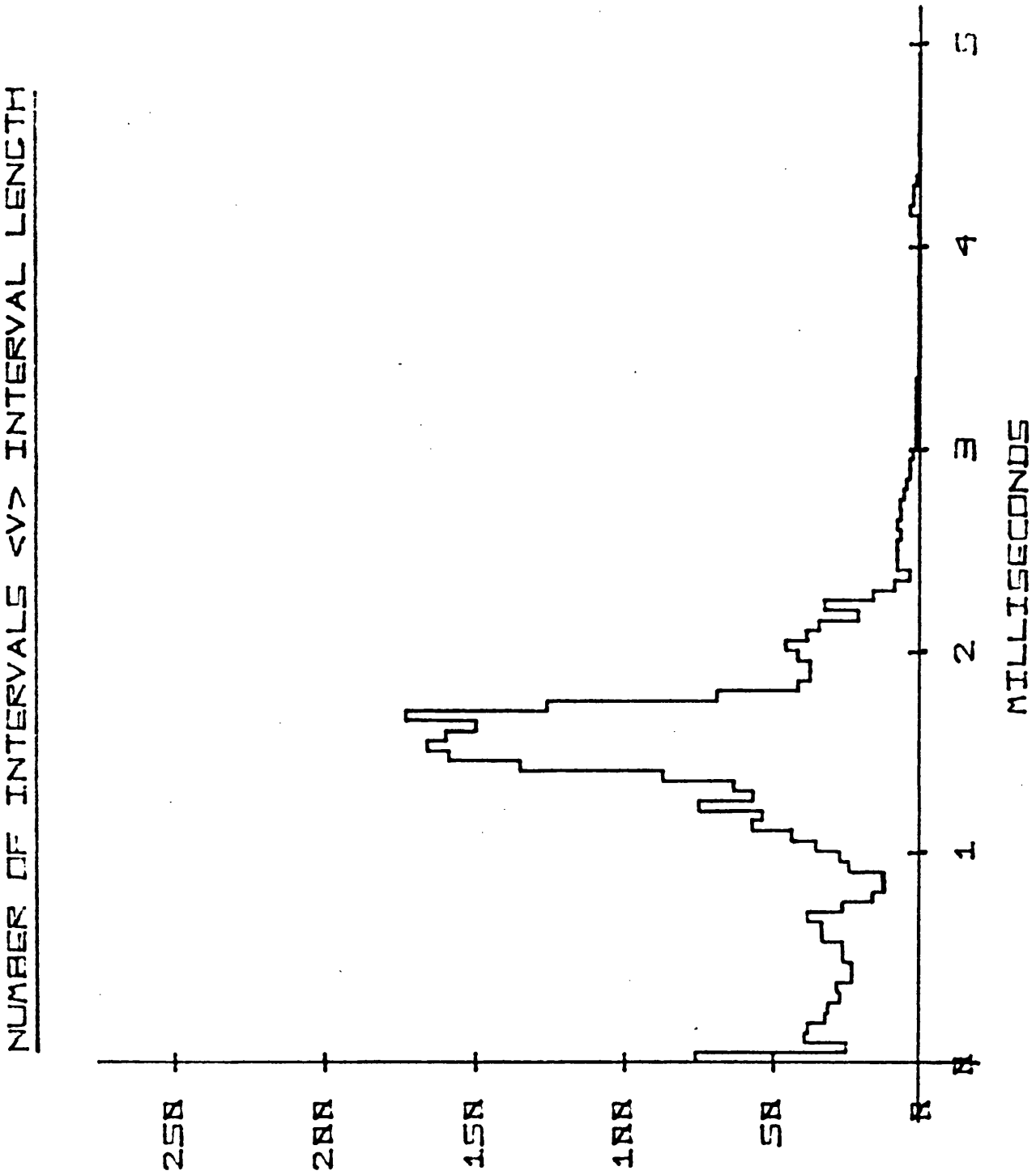


Figure 2.13 The real-zero probability distribution for voiced speech

Threshold : 50mV; Bandwidth : 100-5000Hz.

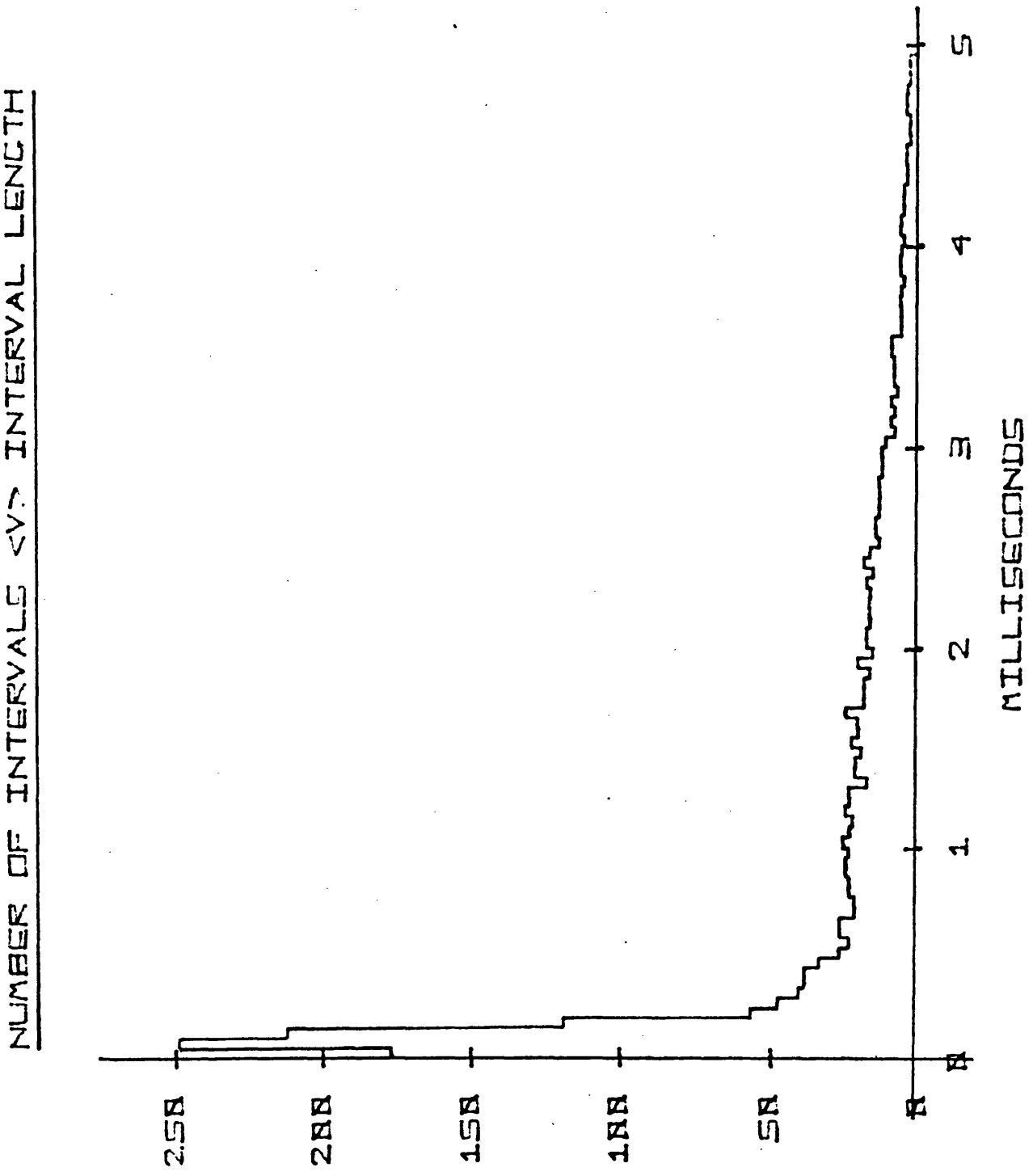


Figure 2.14 The real-zero probability distribution for unvoiced speech

Threshold : 50mV; Bandwidth : 100-5000Hz.

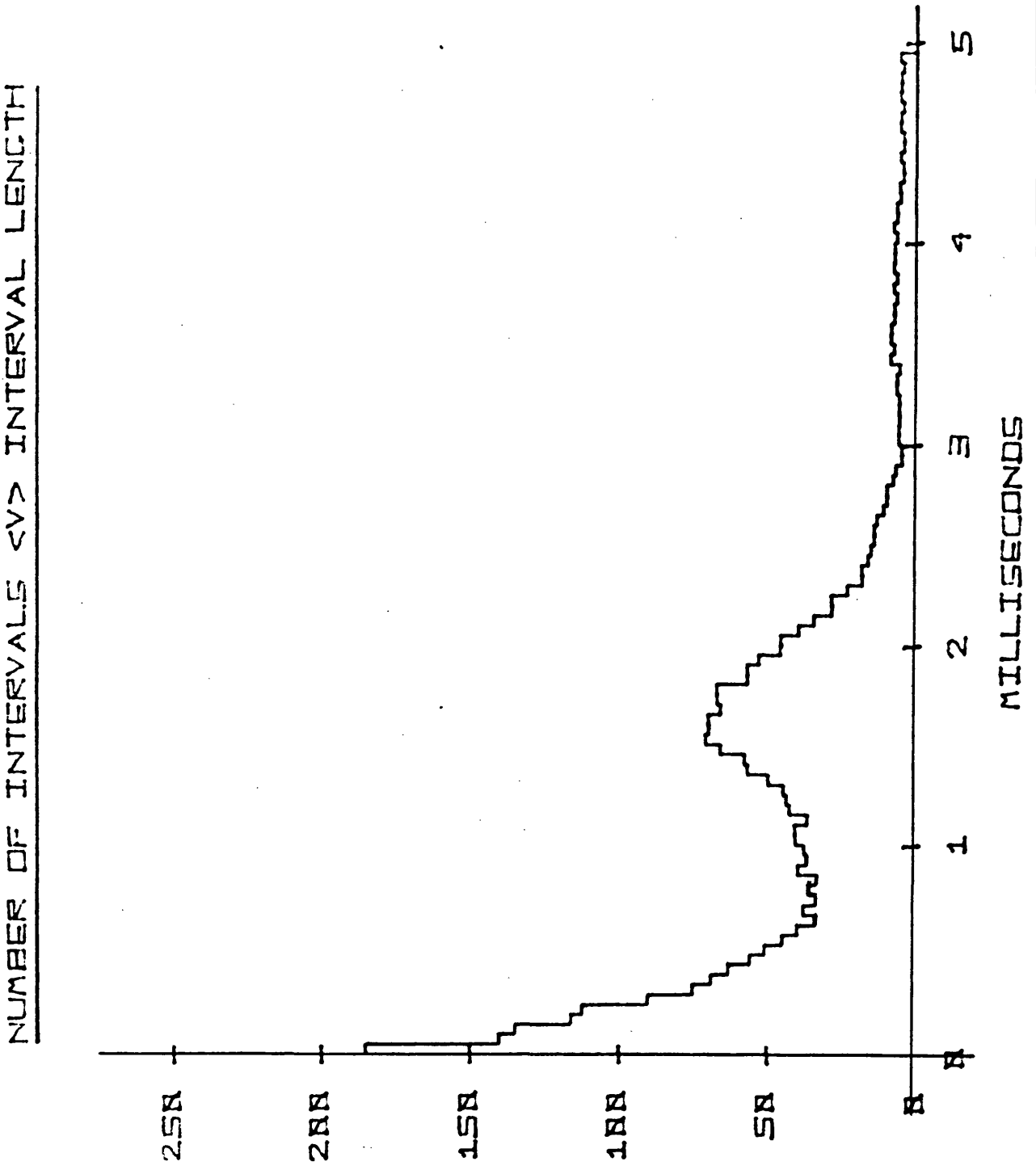


Figure 2.15 The real-zero probability distribution for English speech

Threshold : 25mV; Bandwidth 100-5000Hz.

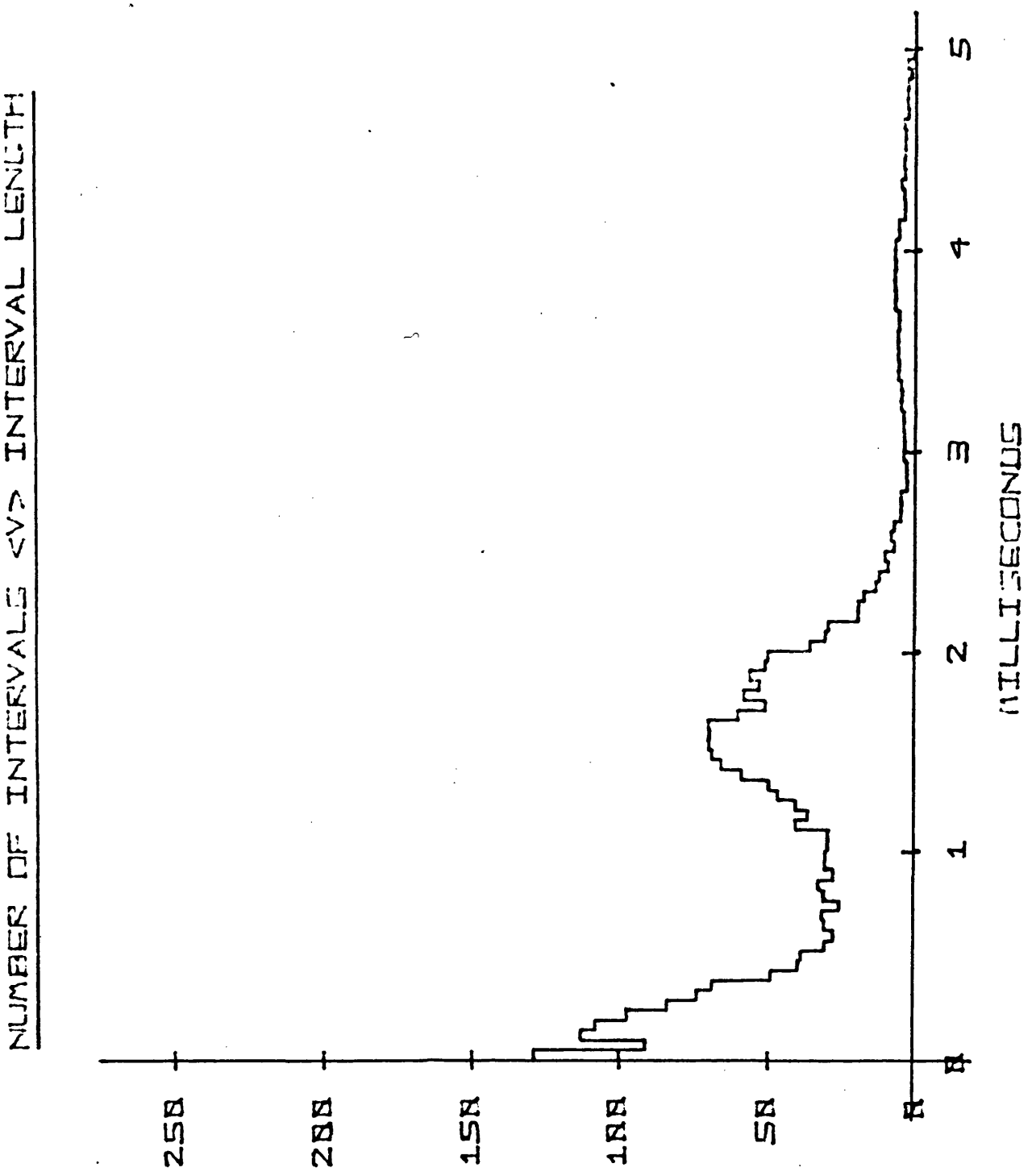


Figure 2.16 The real-zero probability distribution for Punjabi speech

Threshold : 25mV; Bandwidth : 100-5000Hz

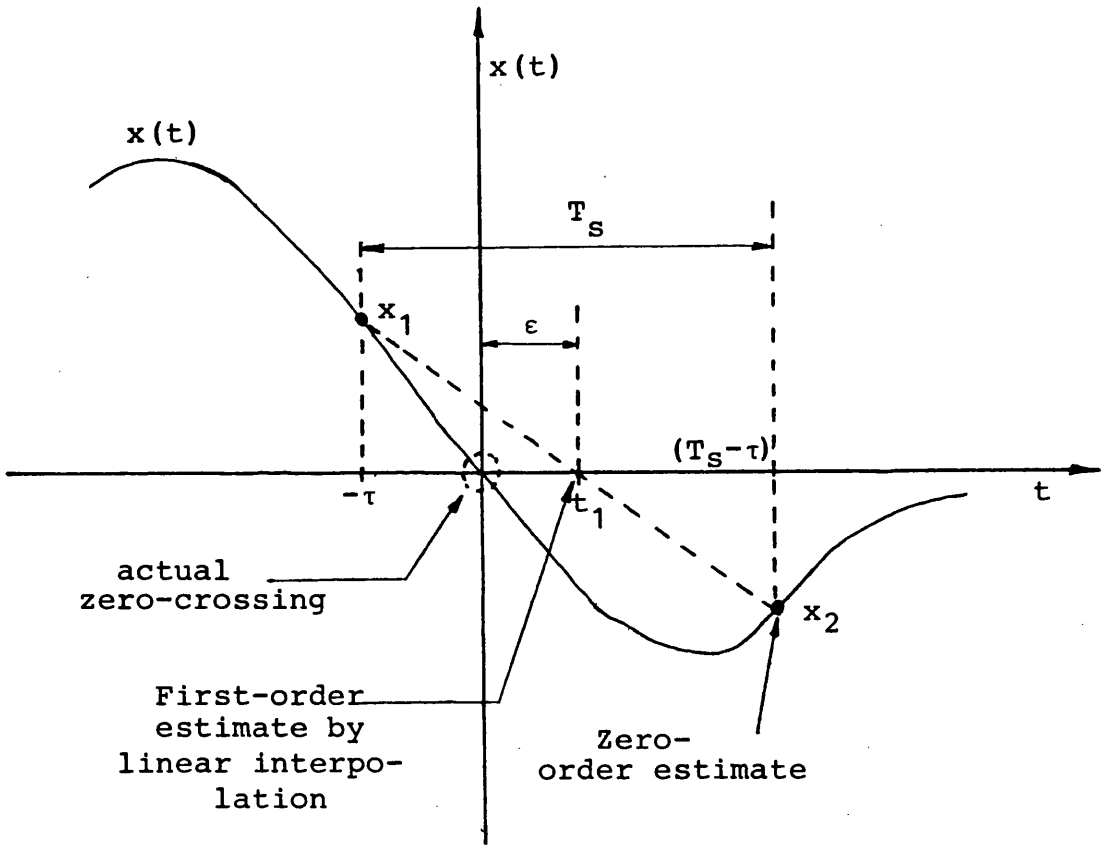


Figure 2.17 Estimation of zero-crossing by interpolation

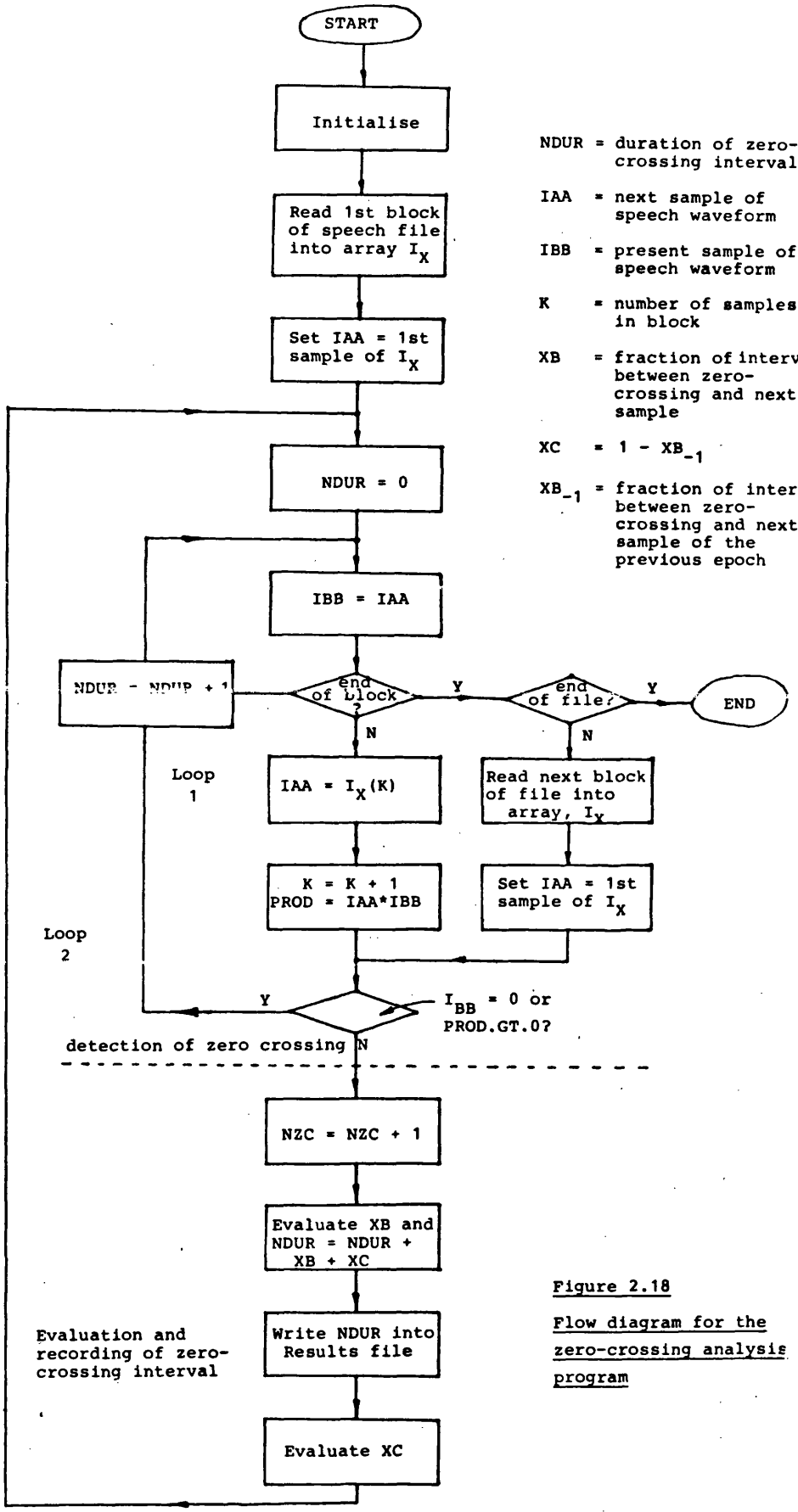


Figure 2.18

Flow diagram for the zero-crossing analysis program

$$y(t) = 100\cos\left(0.11 \times \frac{2\pi t}{T_s}\right) + 150\cos\left(0.055 \times \frac{2\pi t}{T_s}\right)$$

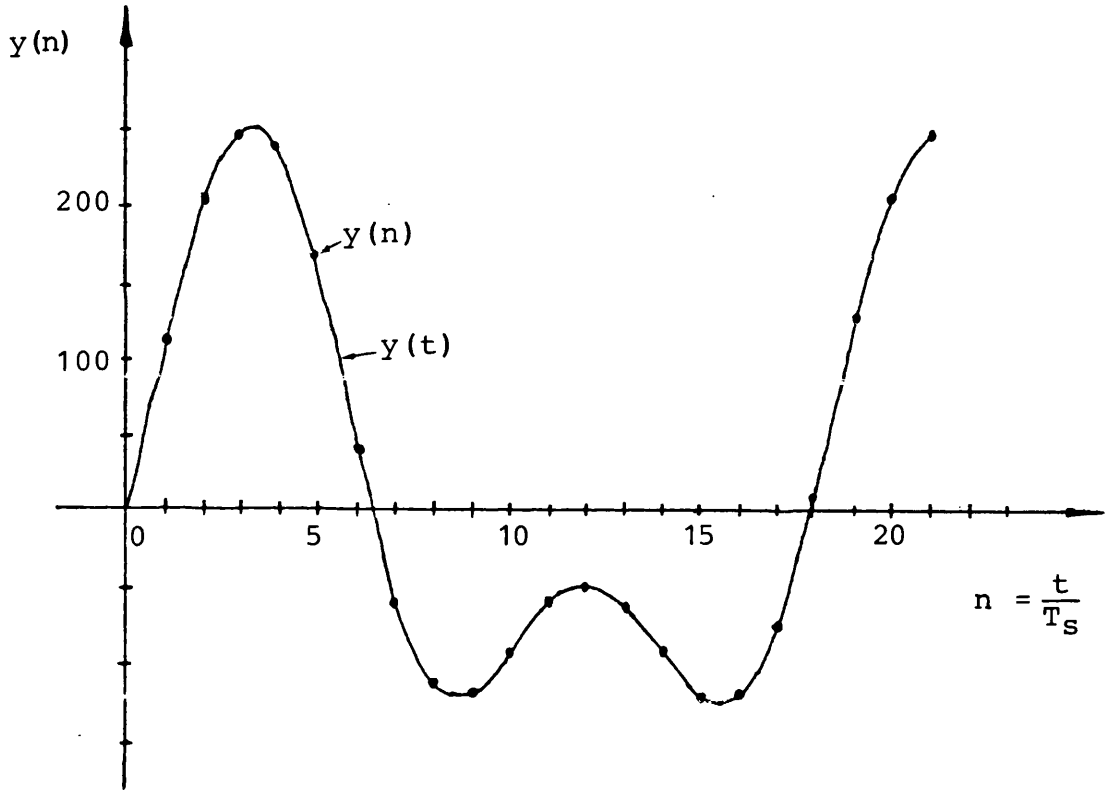


Figure 2.19 Signal used for testing the analysis program

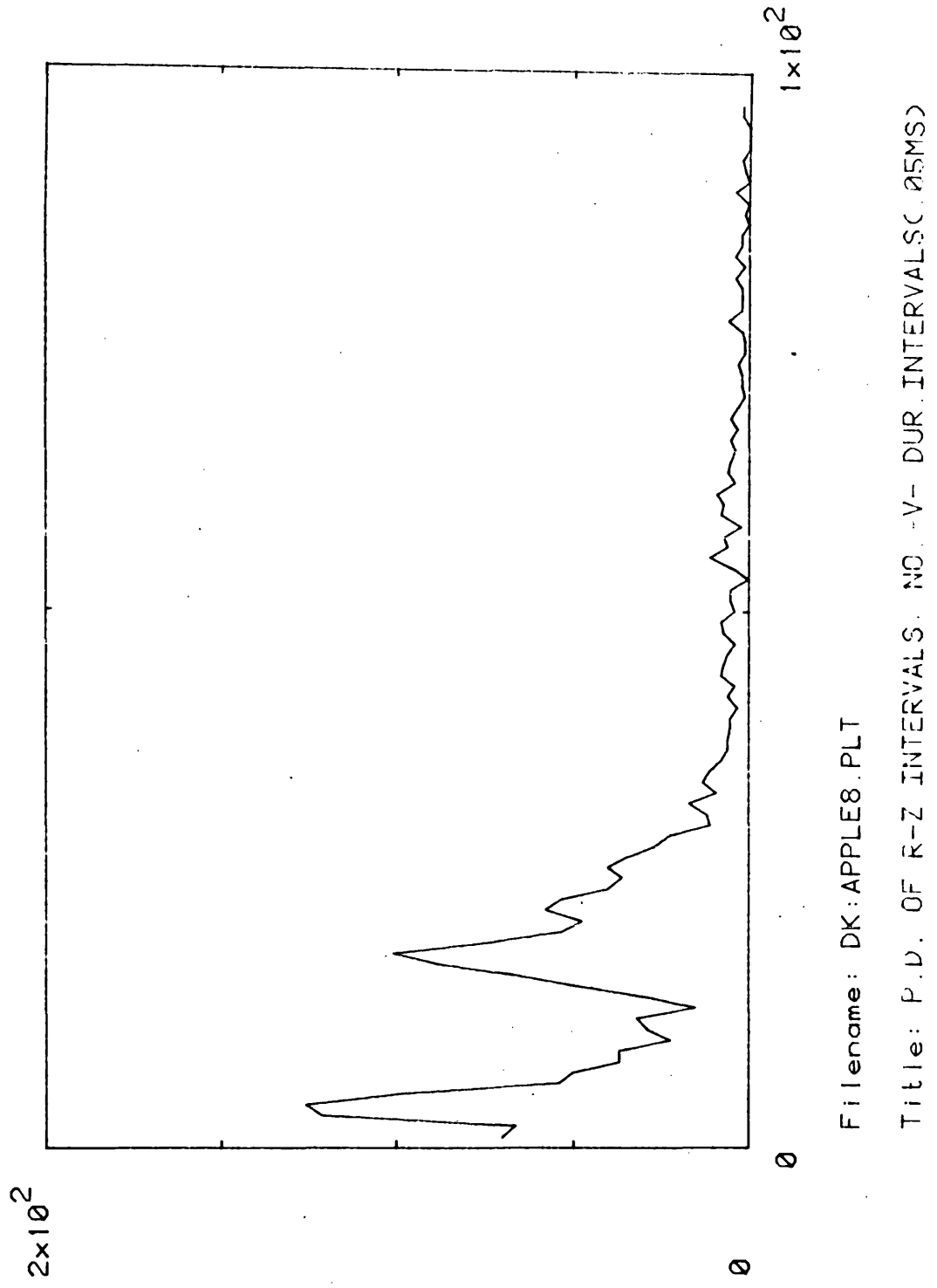


Figure 2.20

Real-zero probability distribution for the speech file APPLE8.SPH

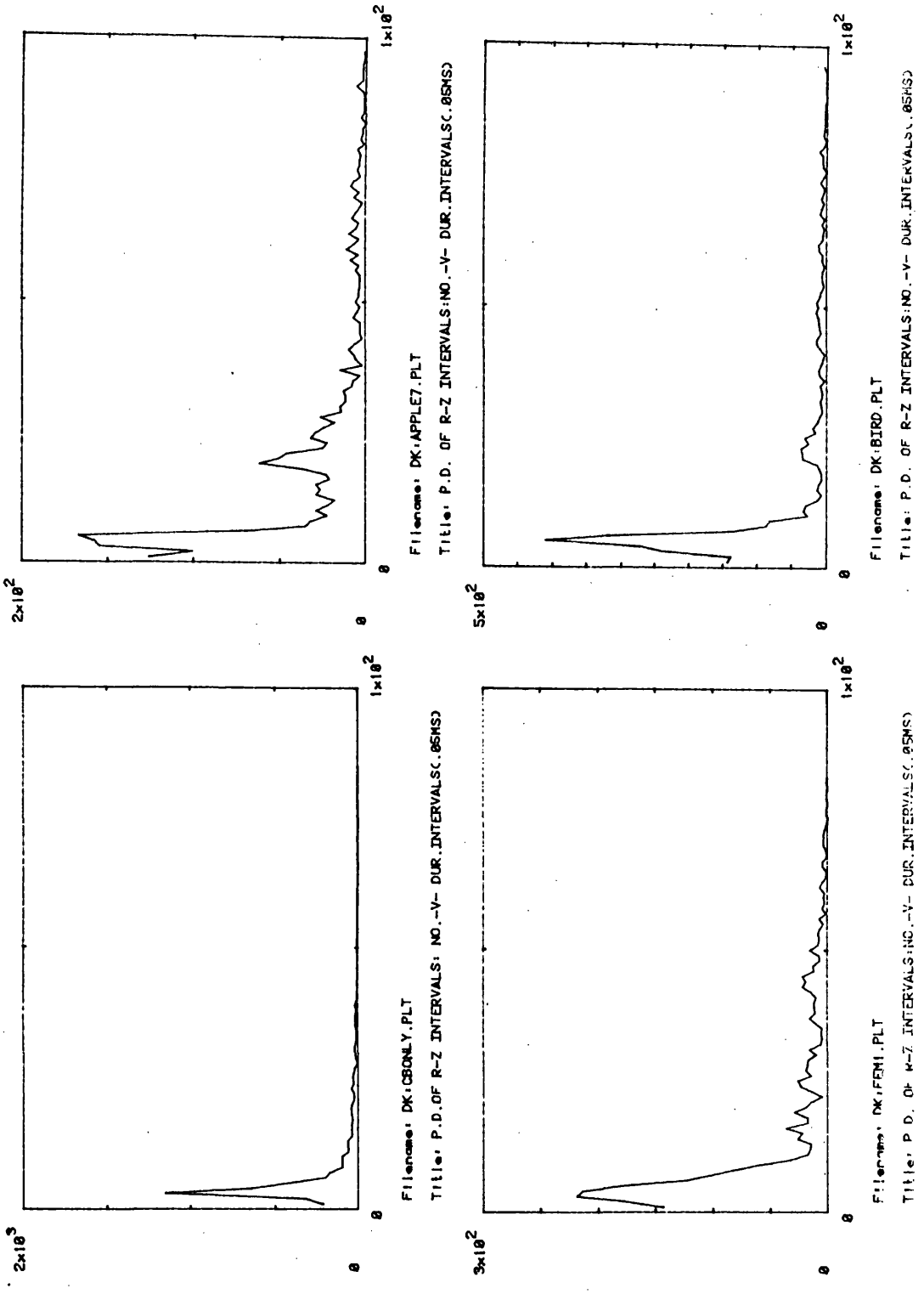


Figure 2.21

The real-zero probability distributions for the speech files CBONLY.SPH, APPLE7.SPH, FEM1.SPH and BIRD.SPH

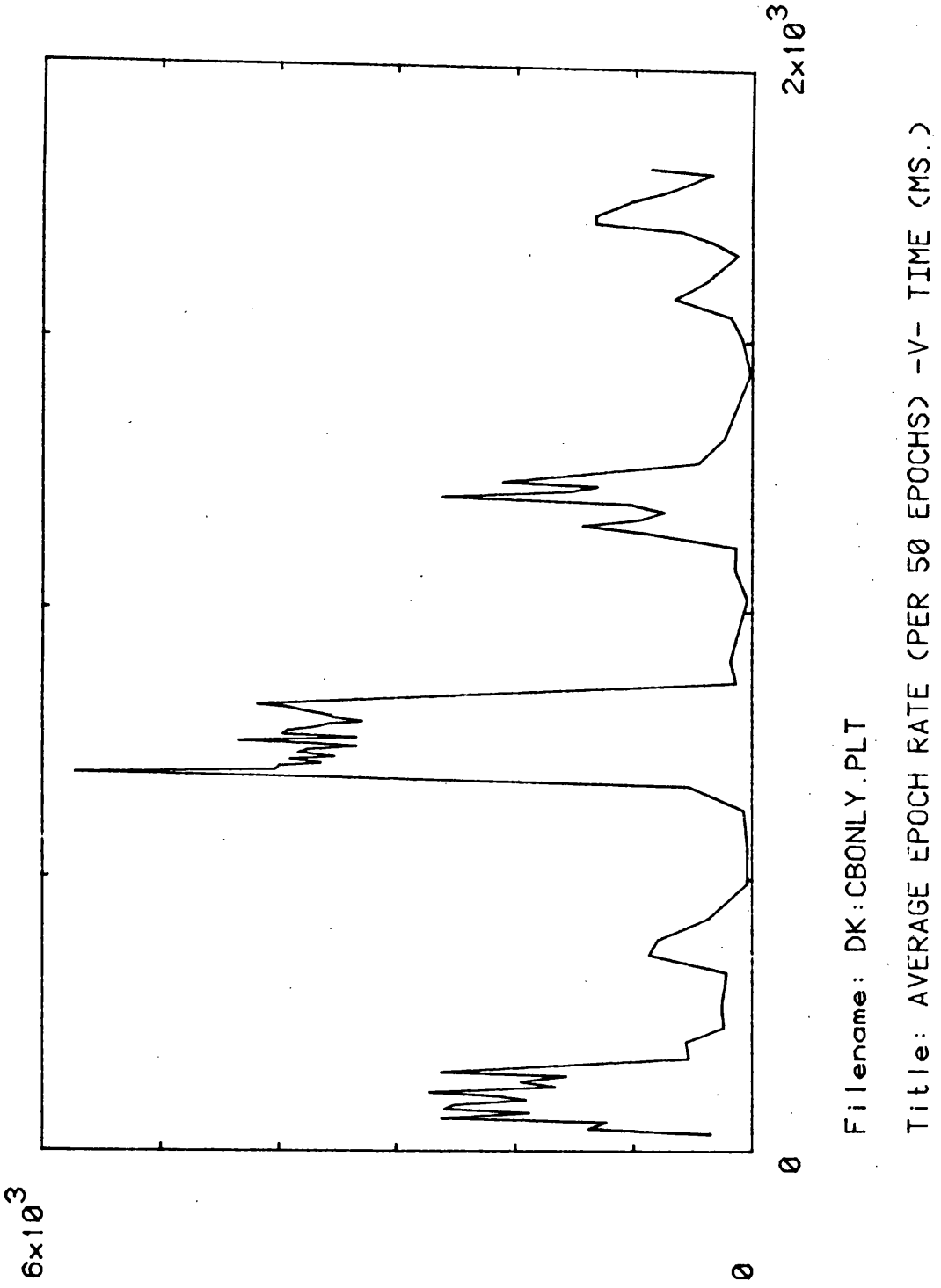


Figure 2.22

The average epoch rate against time for speech file CBONLY.SPH. Illustrates the rise in epoch rate during the unvoiced sounds

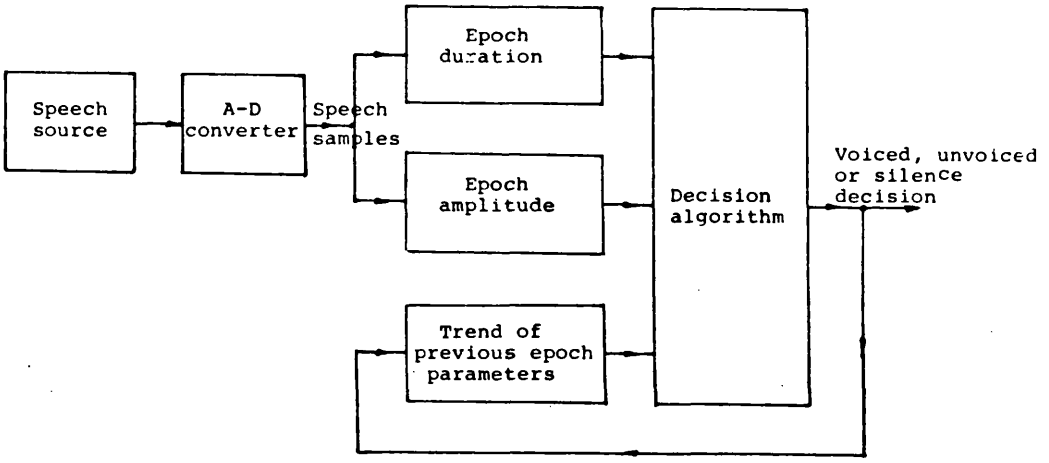


Figure 3.1 Block diagram of the analysis and decision algorithm

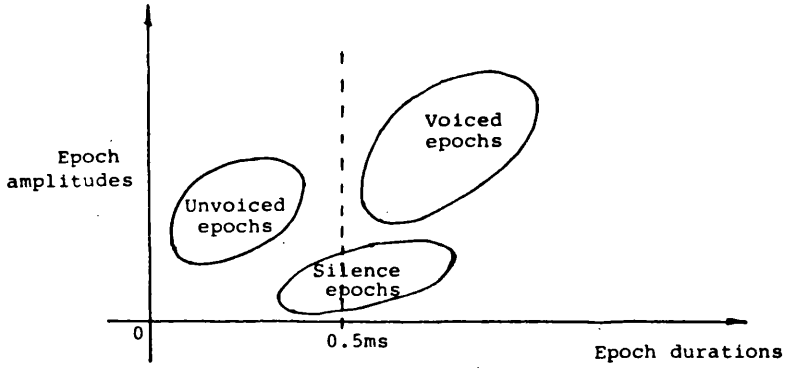


Figure 3.2 Trend chart for speech

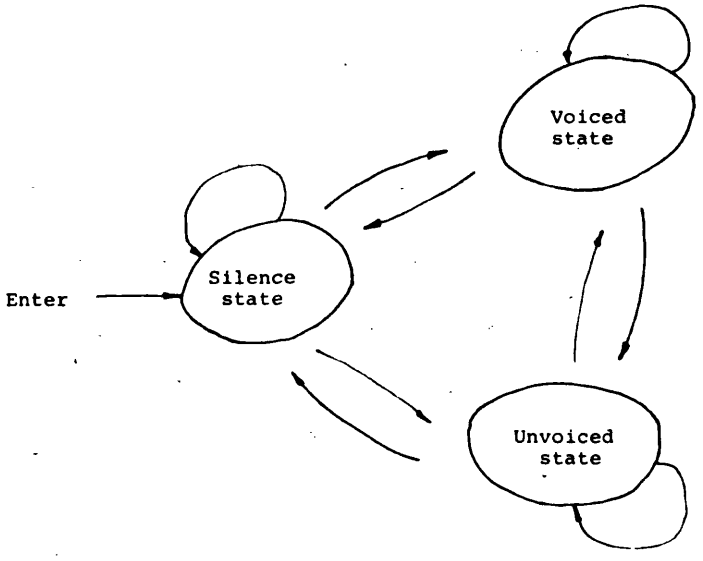


Figure 3.3 State diagram for the decision algorithm

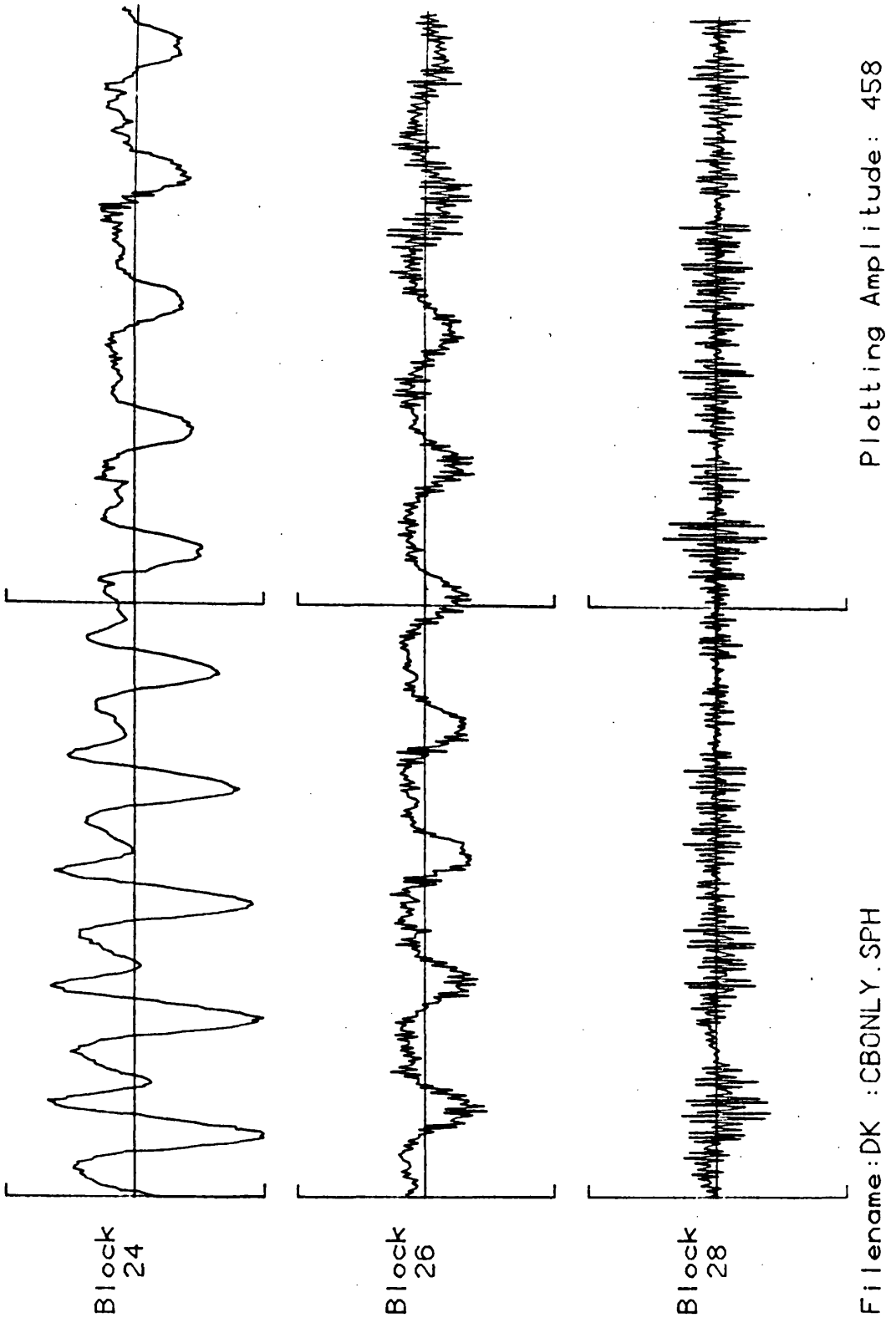


Figure 3.4

Illustrates the transition between the sounds 'l' and 's' in the utterance 'charles' of the speech file CBONLY.SPH

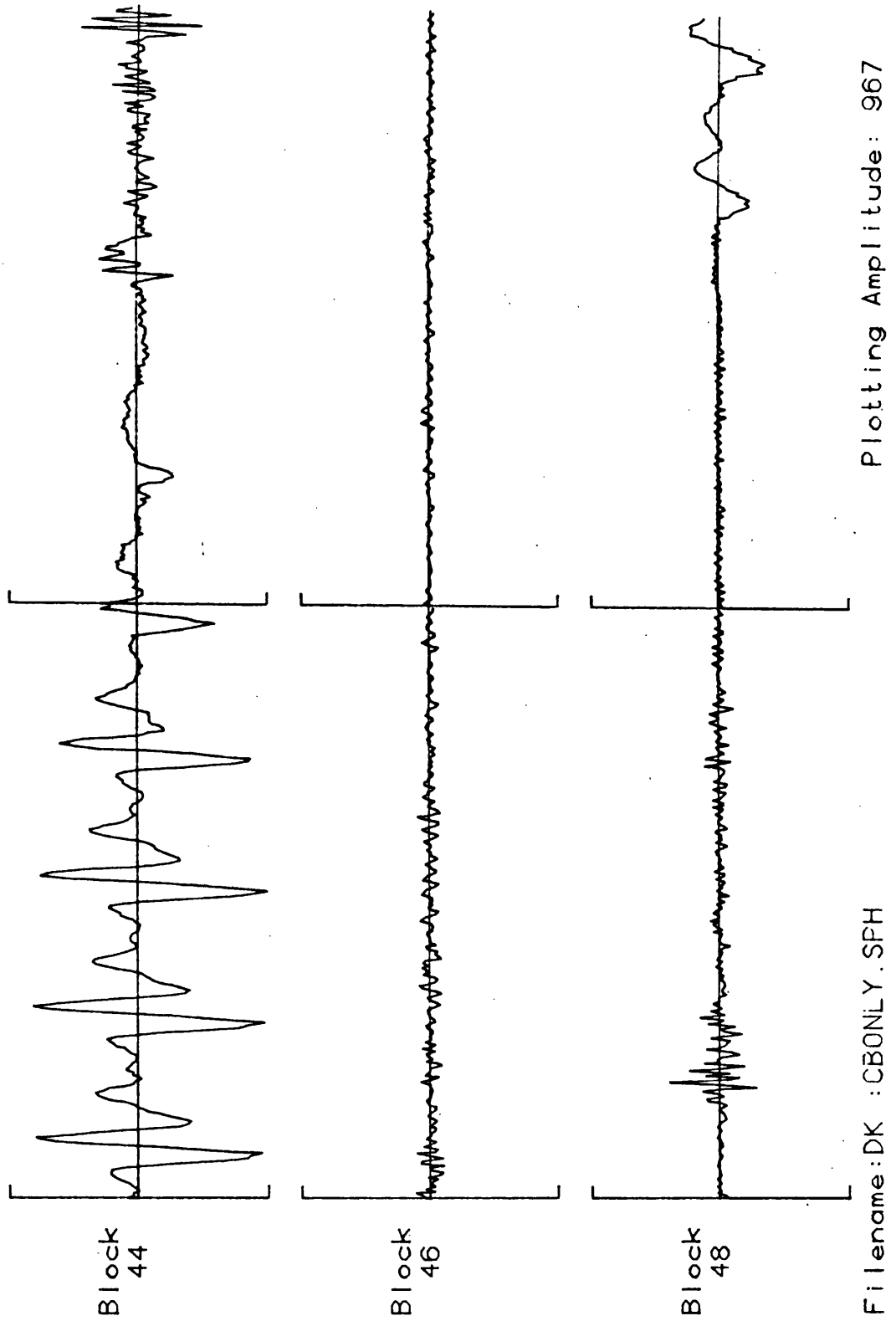


Figure 3.5 Illustrates the transisiton region between the sounds 'o' and 't' in the utterance "bottleneck" (Block 45) of the speech file CBONLY.SPH

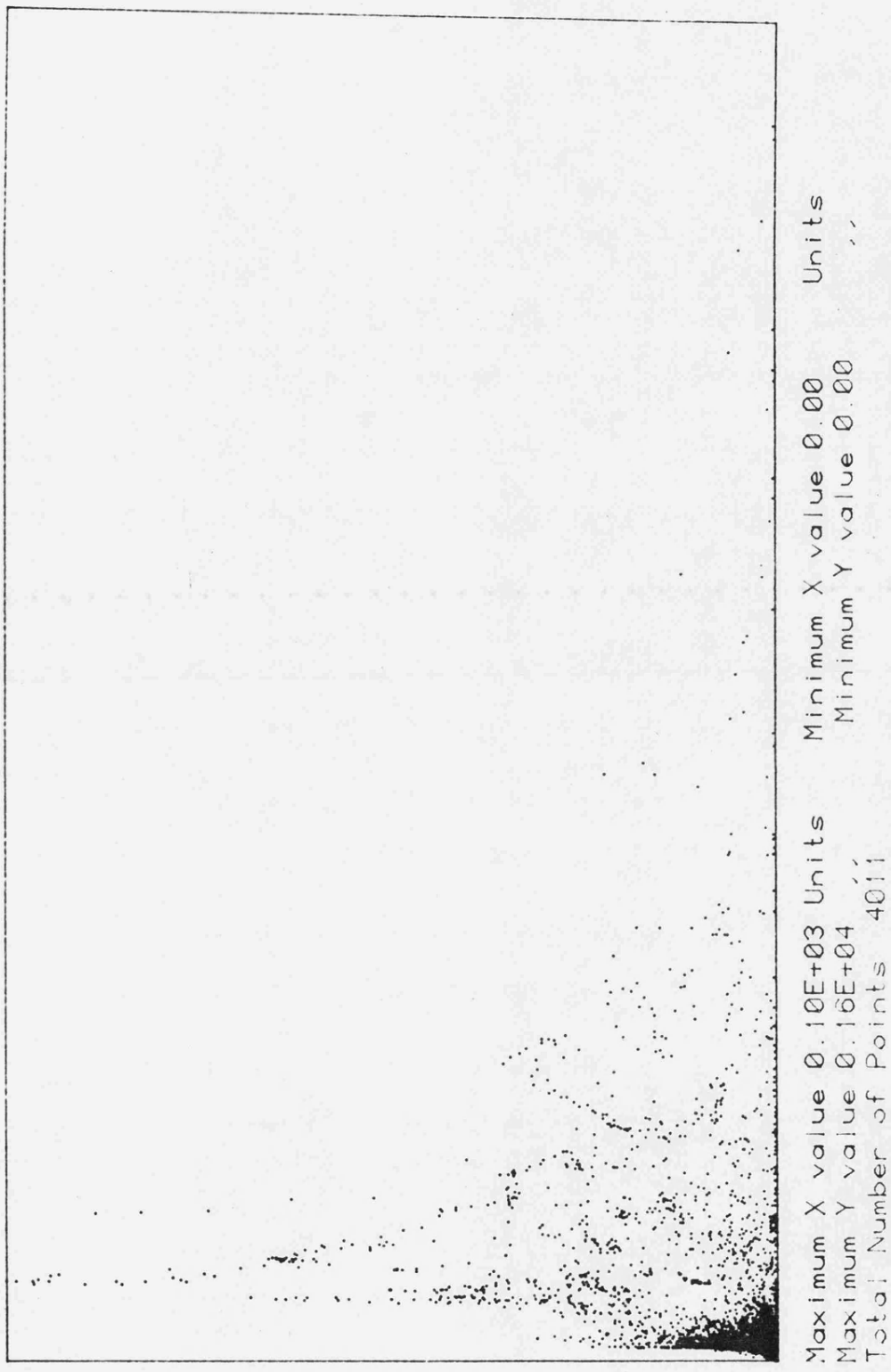


Figure 3.6 Epoch amplitude against epoch duration scatter plot for the speech file CBONLY.SPH

Illustrates the cluster of small amplitude, small duration epochs due to unvoiced speech.

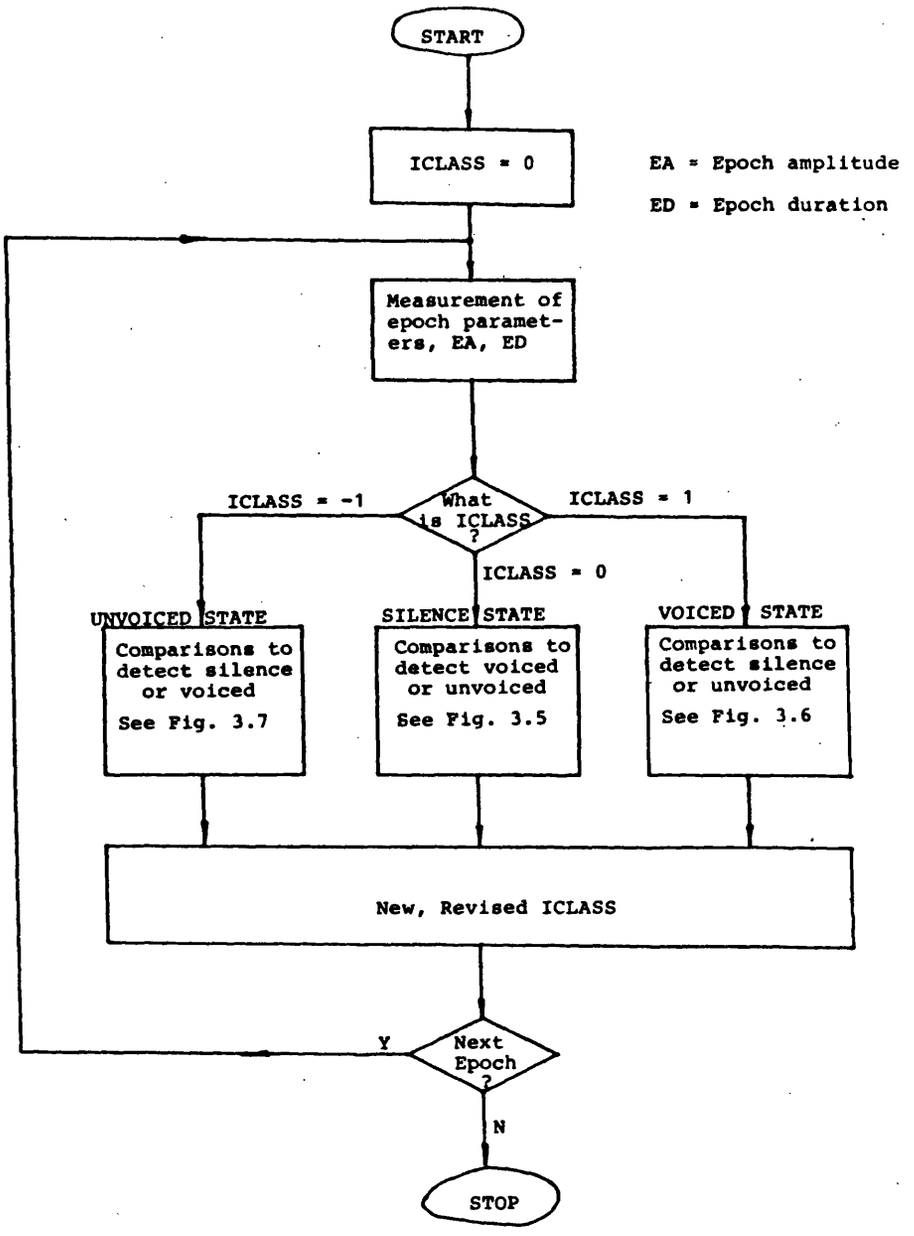


Figure 3.7

Flow Diagram for the Decision Algorithm

ED = Epoch duration
EA = Epoch amplitude
T = Threshold amplitude
KL = Count of consecutive epochs with EA.GT.T
DUR = Voiced/unvoiced duration boundary
ICLASS = Class of speech
PAV = Average epoch amplitude

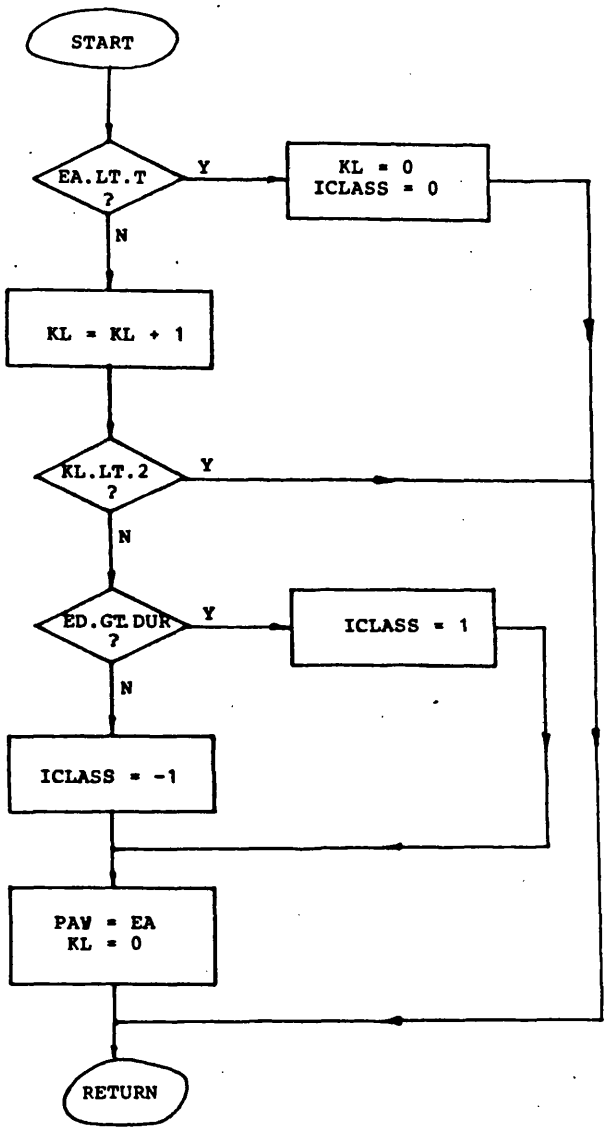


Figure 3.8 Flow diagram for the silence state

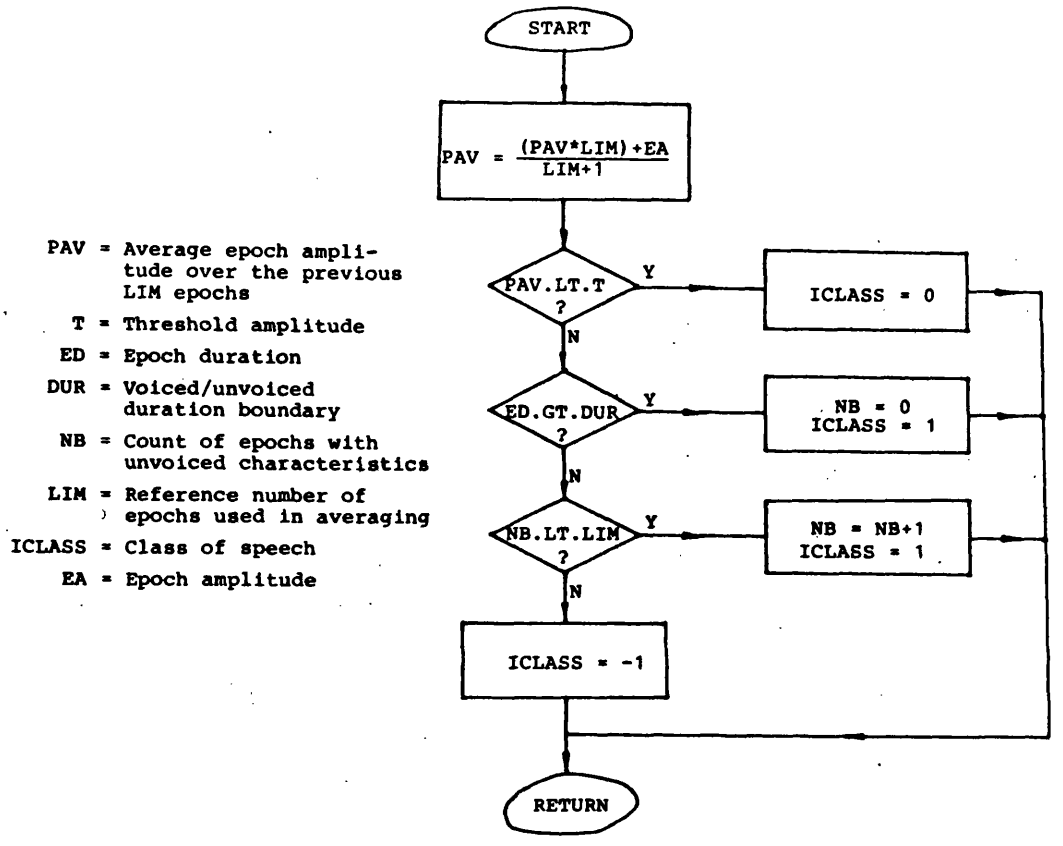


Figure 3.9 Flow diagram for the voiced state

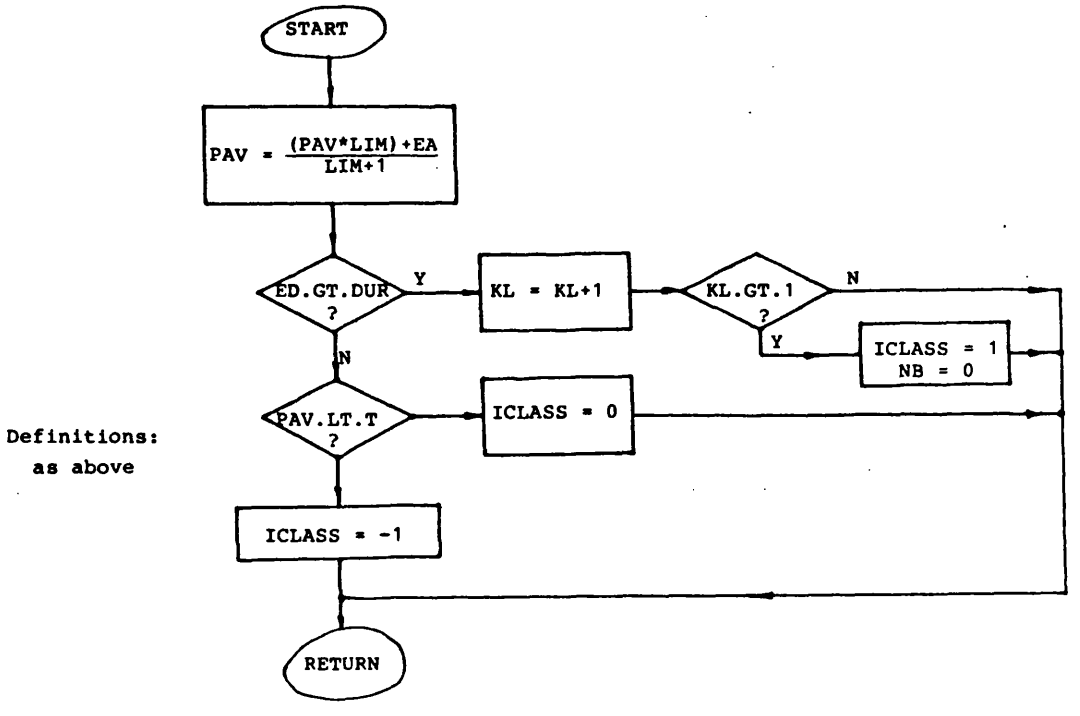


Figure 3.10 Flow diagram for the unvoiced state

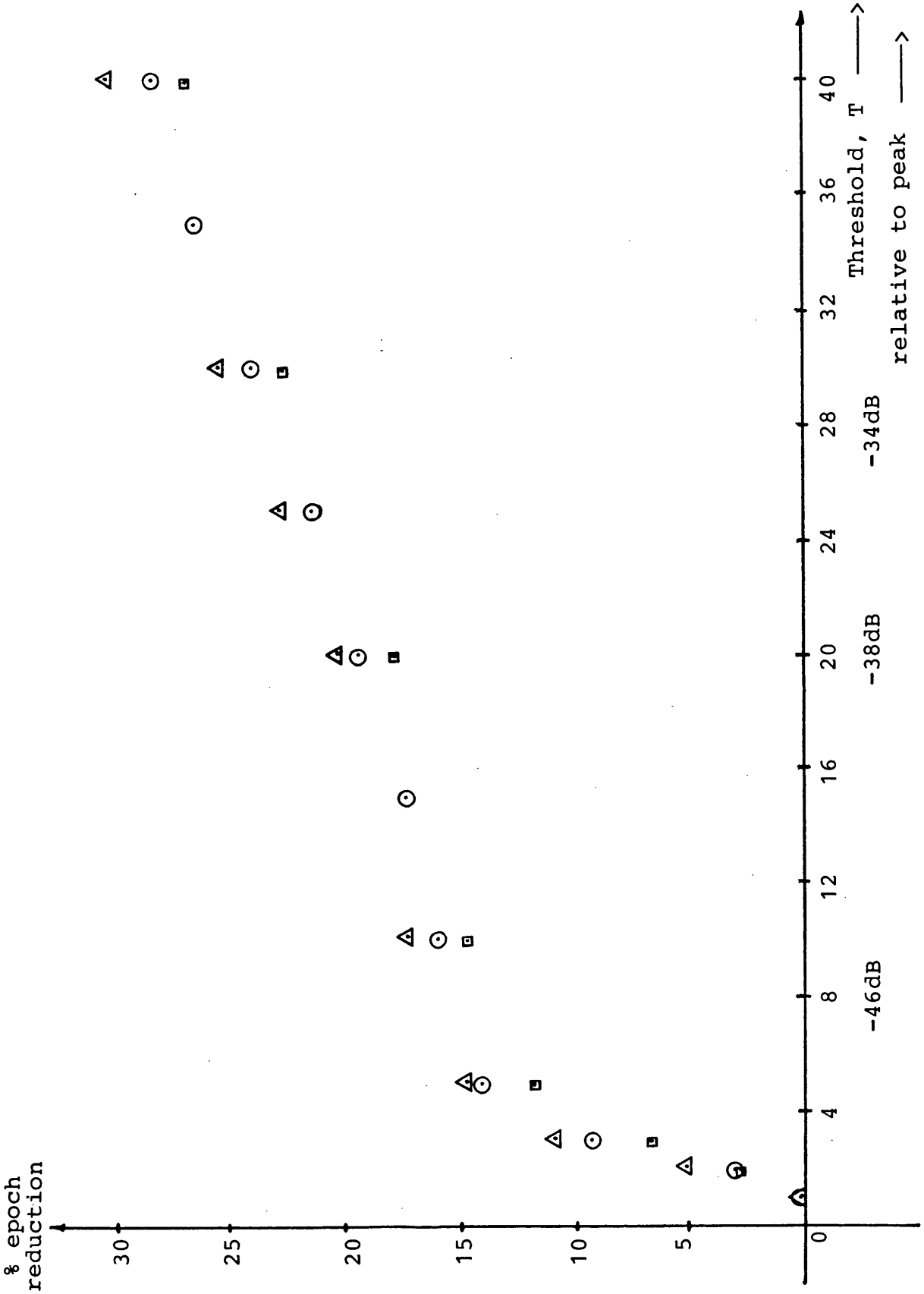


Figure 3.11 Epoch reduction with varying threshold values for the speech file CBONLY.SPH

△ Dur : 40 Lim : 20

⊙ Dur : 40 Lim : 30

⊡ Dur : 40 Lim : 40

Epoch Count with Threshold = 20.0		
LIM	DUR	
	3.0	4.0
15	1334	1315
20	1385	1344
25	1399	1357
30	1430	1389
35	1514	1475
40	1559	1519

Figure 3.12 Epoch Count with Different Values of LIM and DUR for Speech File CBONLY

Checking for Distortion: D = distortion Threshold = 20.0 ✓ = no distortion		
LIM	DUR	
	3.0	4.0
5	D	D
10	D	D
15	D	D
20	D	D
25	✓	✓
30	✓	✓
35	✓	✓
40	✓	✓
45	✓	✓

Figure 3.13 Distortion with Different Values of LIM and DUR for Speech File CBONLY

Checking for sensivity of change of class. Table indicates block numbers of speech file.

LIM	DUR					
	1.0	2.0	3.0	4.0	5.0	6.0
5	Not enough information to critically decide on class					
10						
15	Not enough information to critically decide on class, ie. calls unvoiced-voiced		2,45,47,49 67,68,69	2,47,49,64 65,69	Replaces epochs within voiced region	
20			2,47,49,69	2,47,49,69		
25			2, 69	2, 69		
30			69	69		
35			69	69		
40			-	-		
45						
50	Too little epoch reduction					
55						
60						

Figure 3.14 Change of Class Sensitivity with Different Values of LIM and DUR for Speech File CBONLY

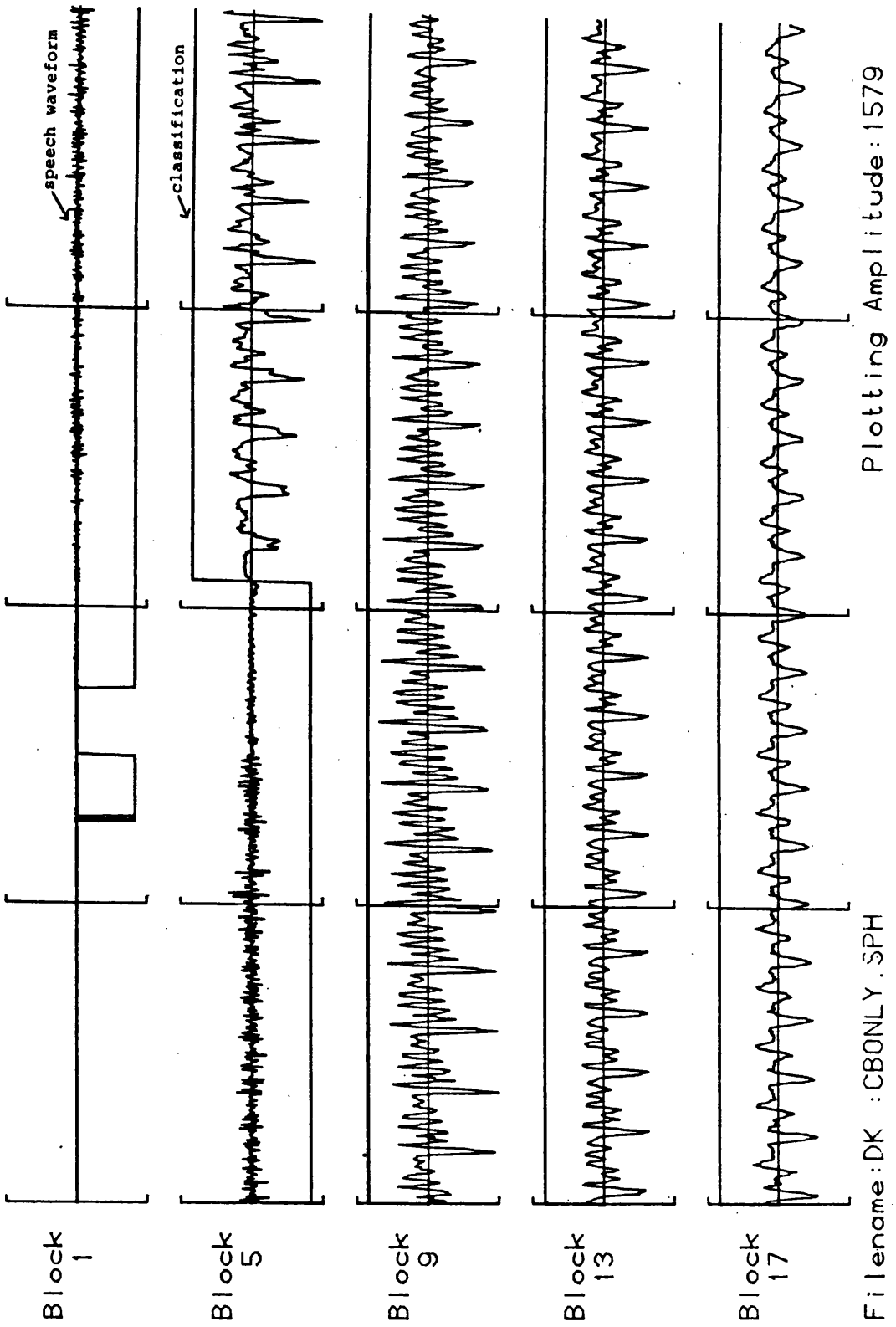


Figure 3.15(a)

Classification of the speech file CBONLY.SPH

Positive level indicates voiced speech
Zero level indicates silence
Negative level indicates unvoiced speech

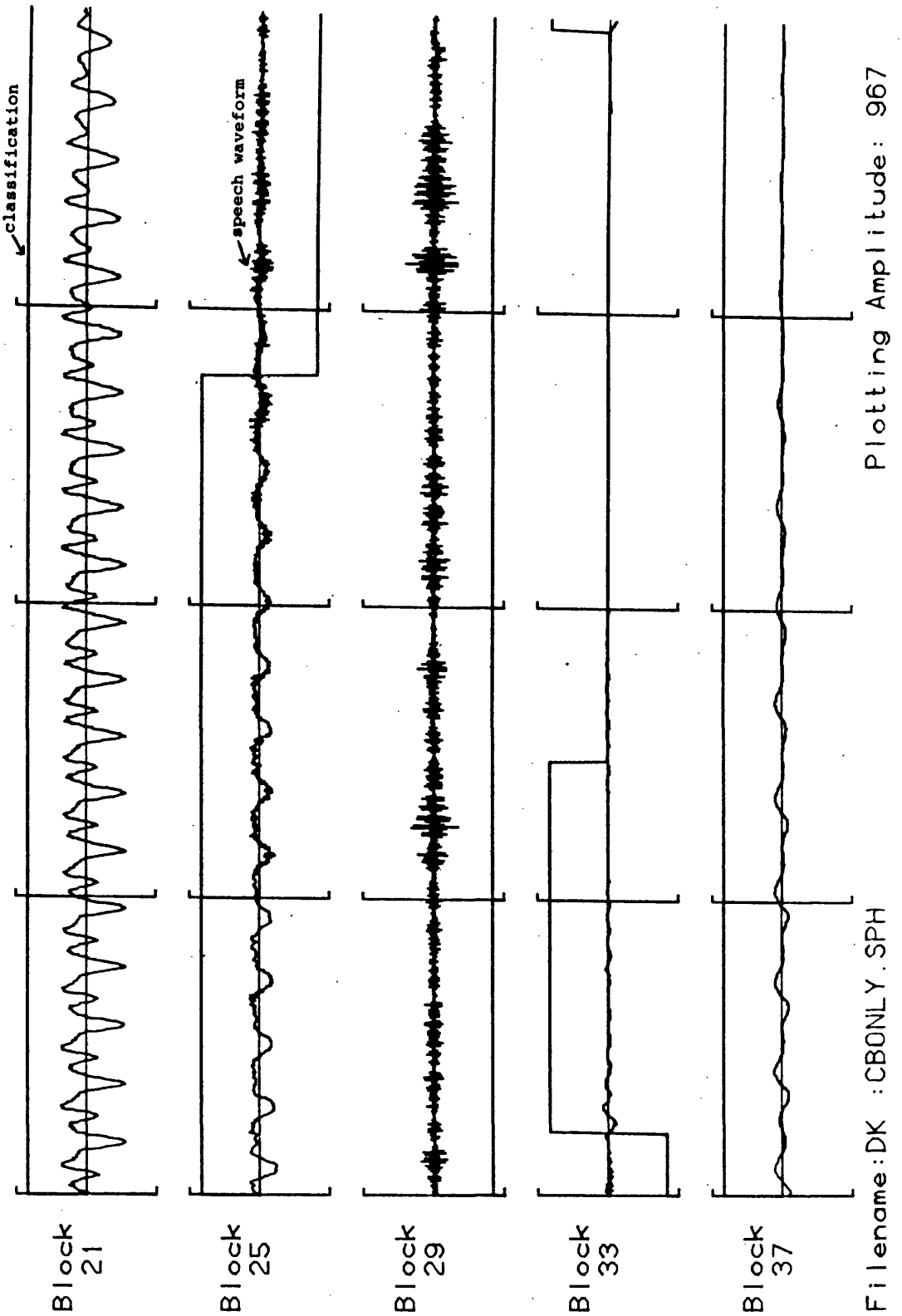


Figure 3.15(b)

Classification of the speech file CBONLY.SPH

Positive level indicates voiced speech
Zero level indicates silence
Negative level indicates unvoiced speech

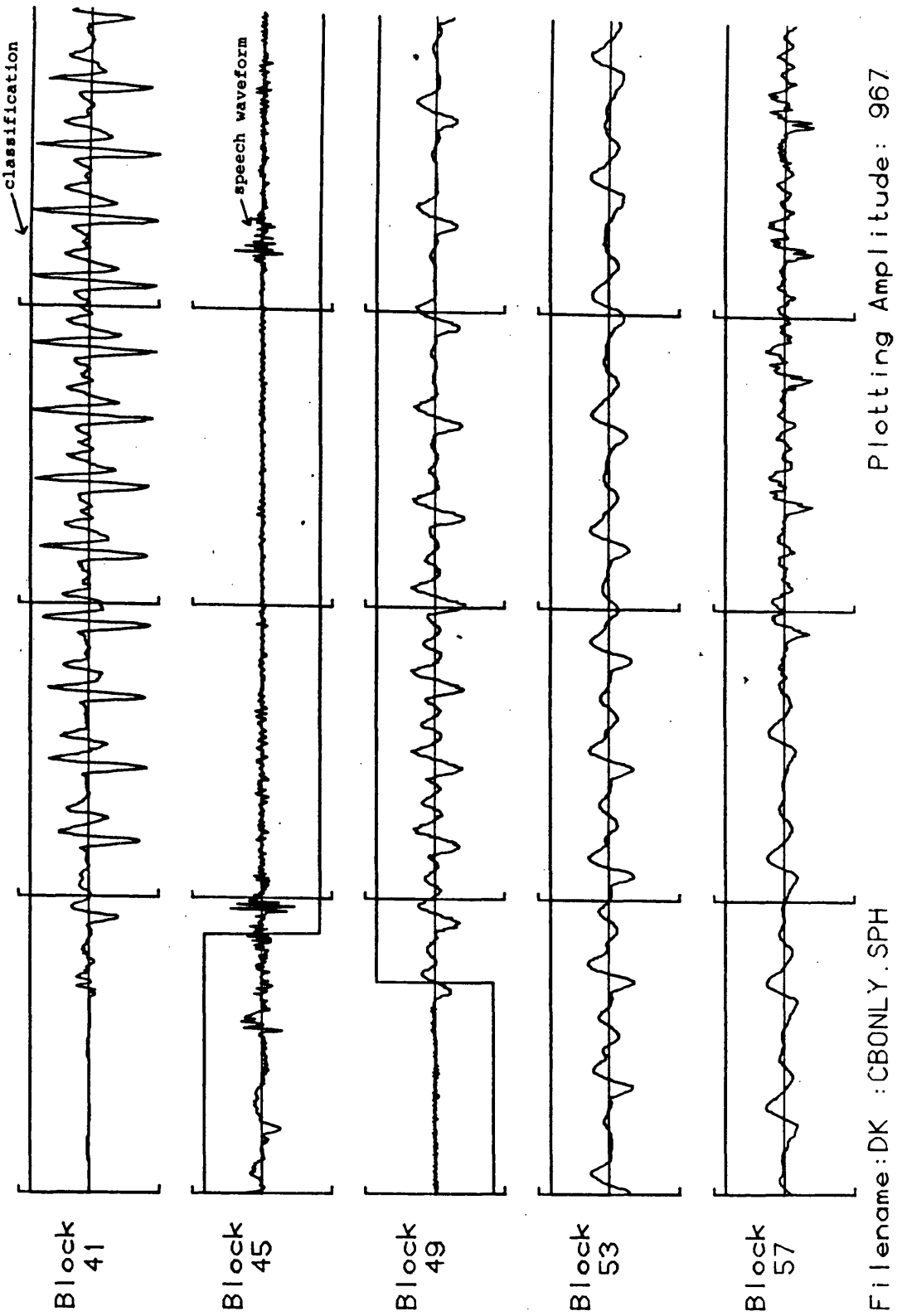


Figure 3.15(c)

Classification of the speech file CBONLY.SPH

Positive level indicates voiced speech
Zero level indicates silence
Negative level indicates unvoiced speech

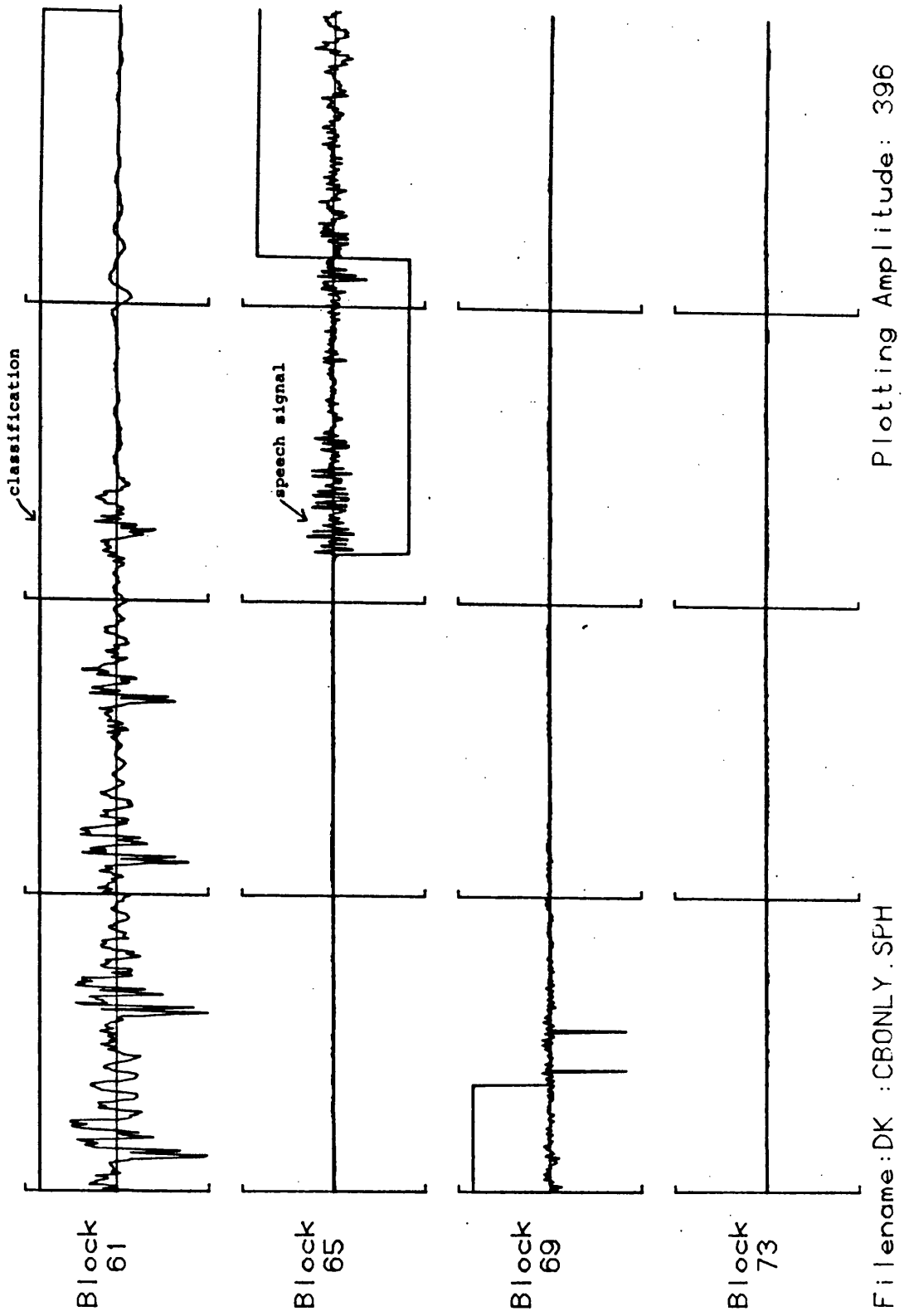


Figure 3.15(d)

Classification fo the speech file CBONLY.SPH

Positive level indicates voiced speech
Zero level indicates silence
Negative level indicates unvoiced speech

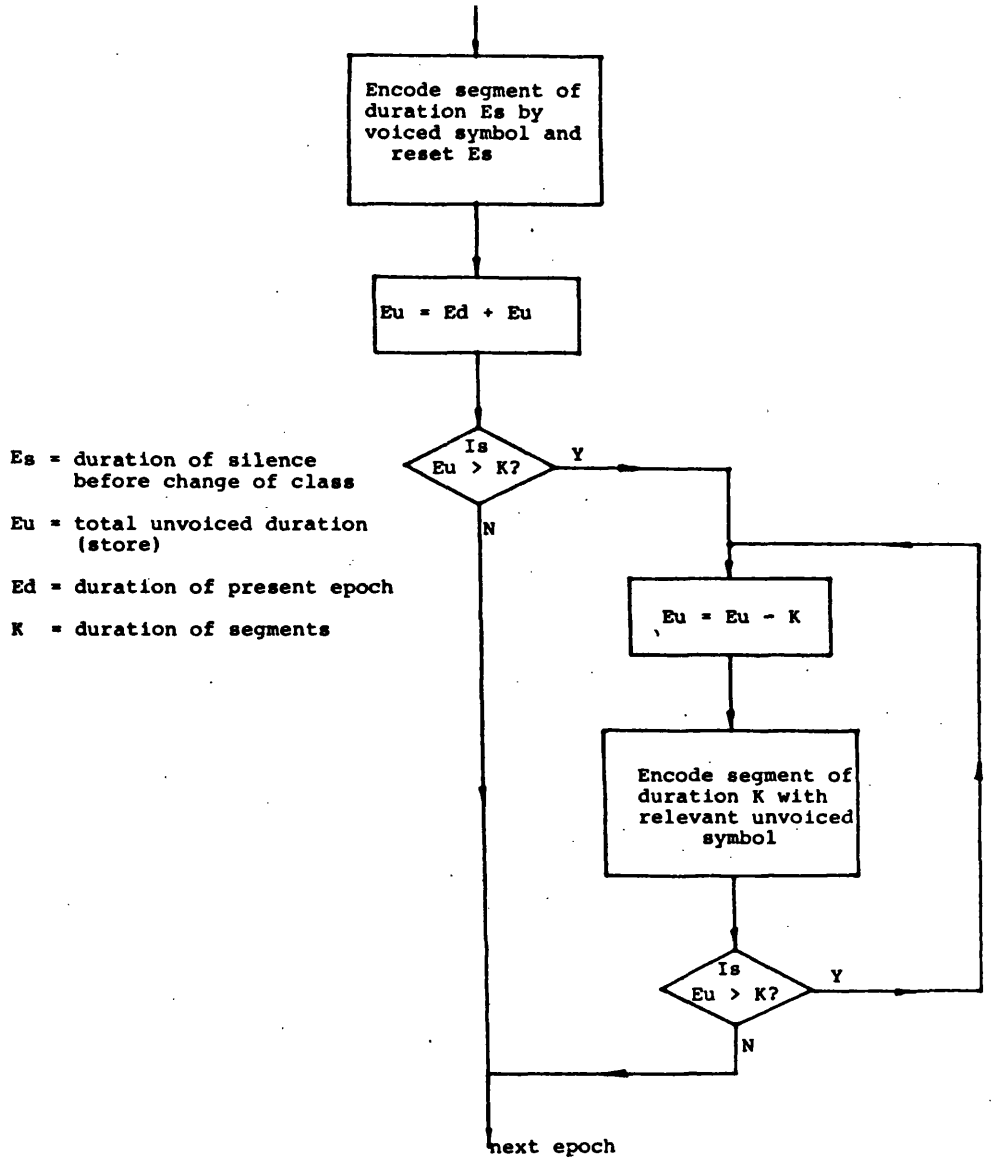


Figure 4.1

Flow diagram for segmenting unvoiced regions

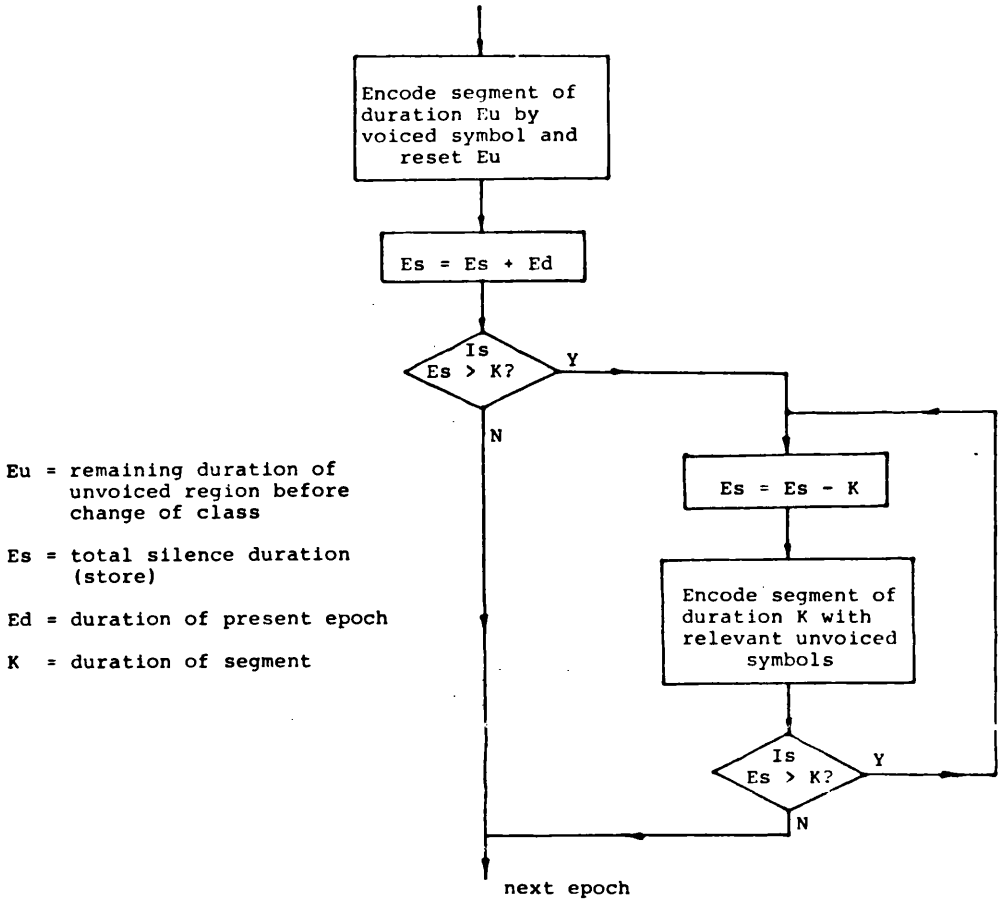


Figure 4.2

Flow diagram for segmenting silence regions

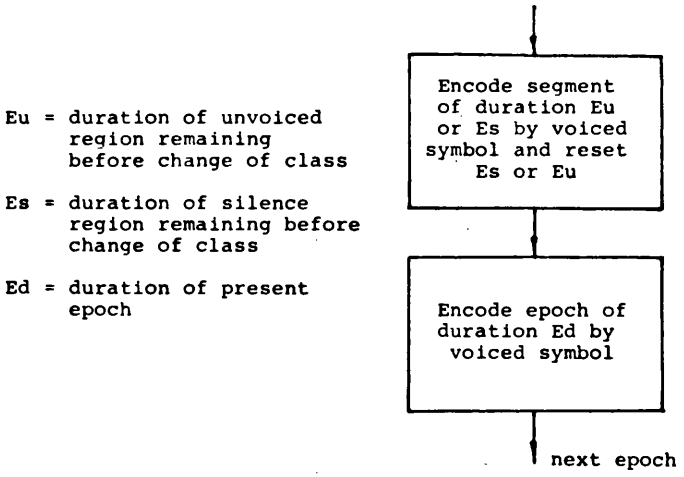


Figure 4.3

Flow diagram for voiced regions

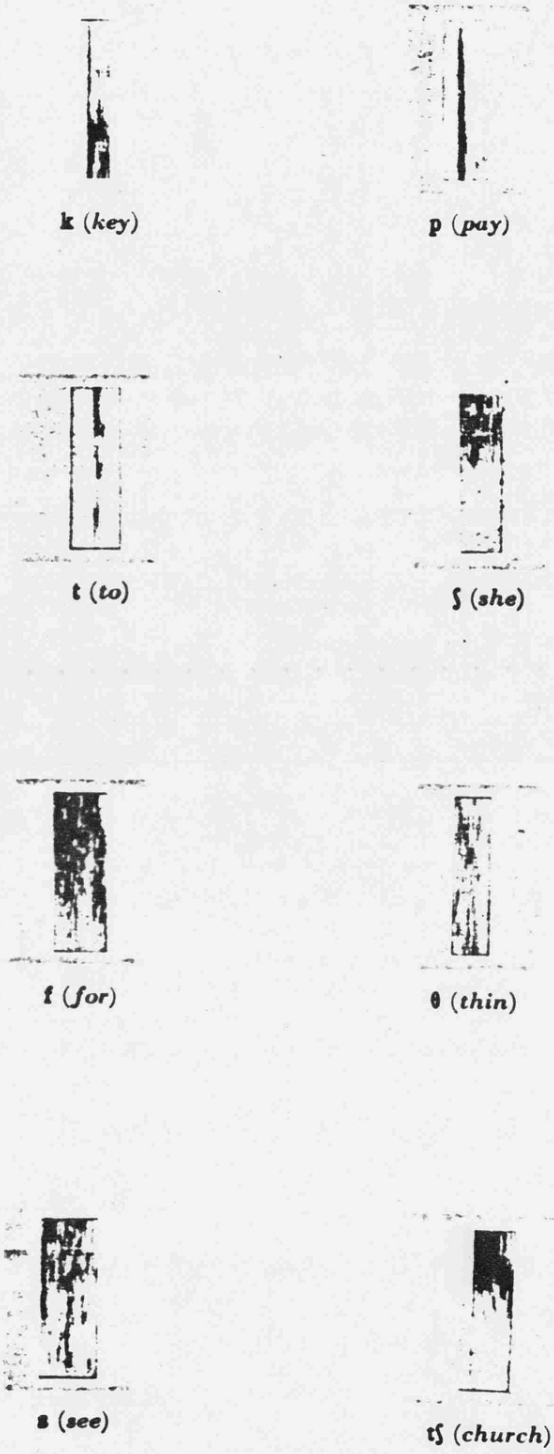


Figure 4.4 Spectrograms of the different unvoiced sounds (from Potter et al)

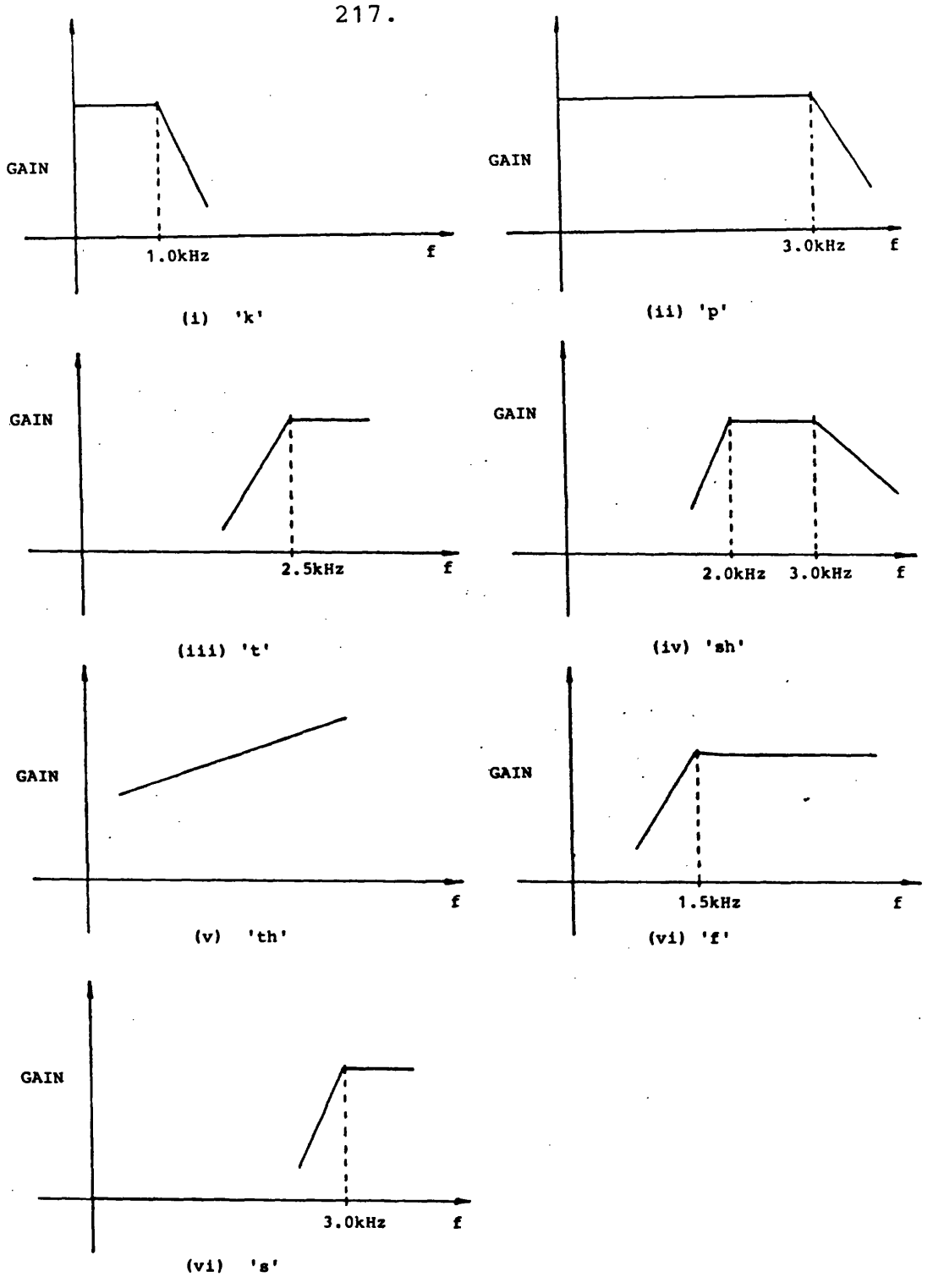


Figure 4.5 Estimation of the noise filter gain responses required to produce the above sounds

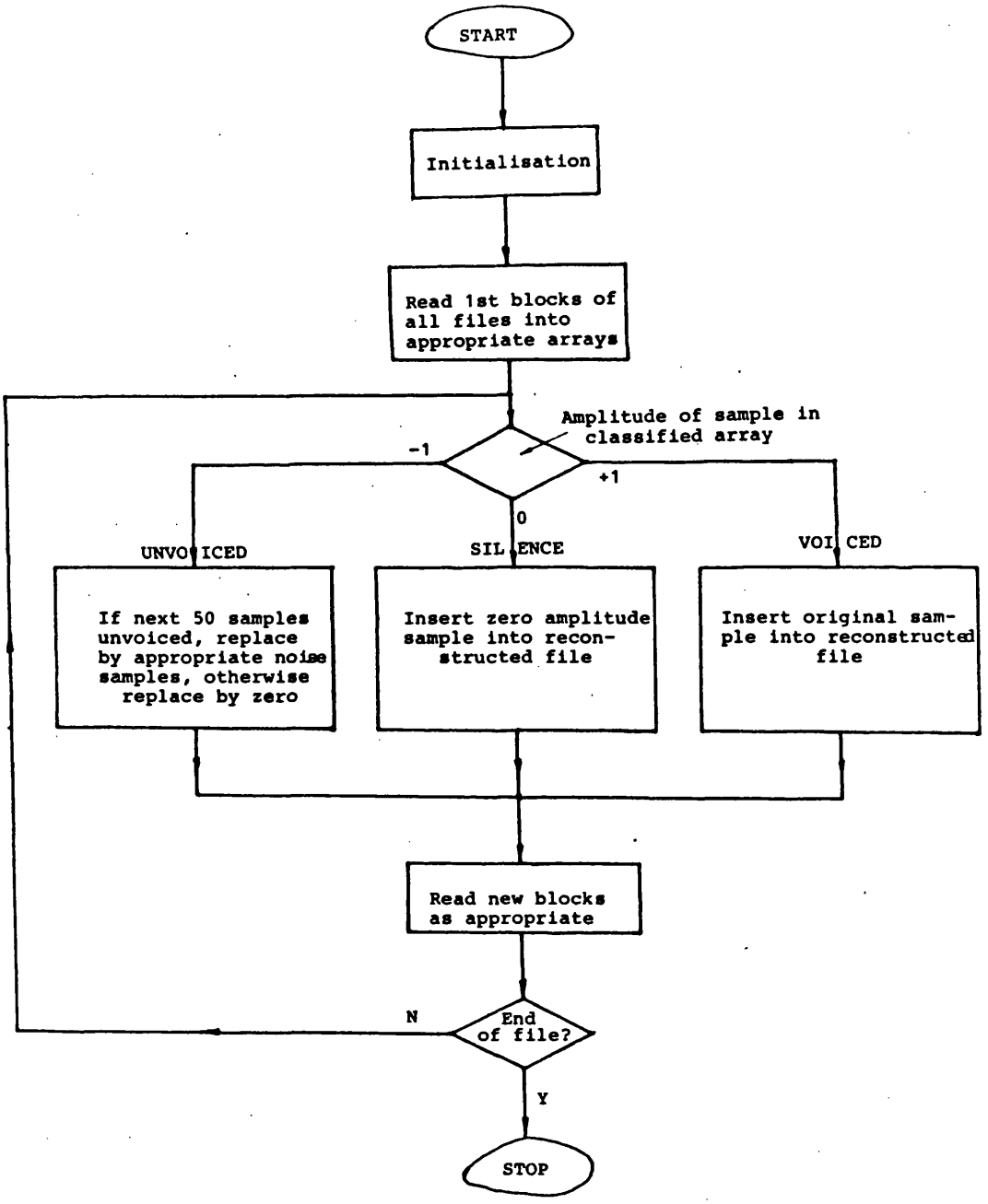


Figure 4.6

Flow diagram for RECON9

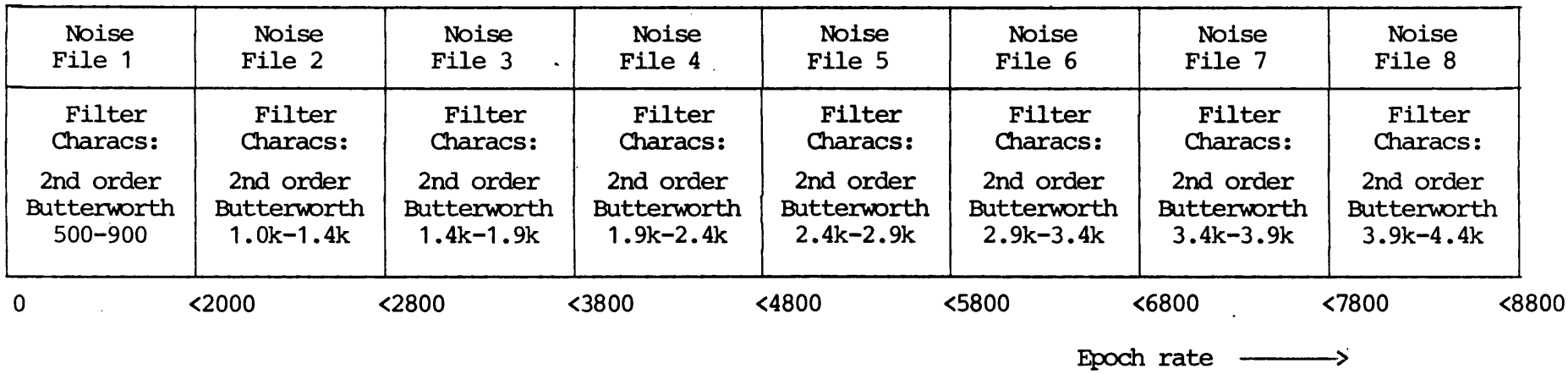


Figure 4.7 Subdivision of epoch rate variation range between 8 noise files

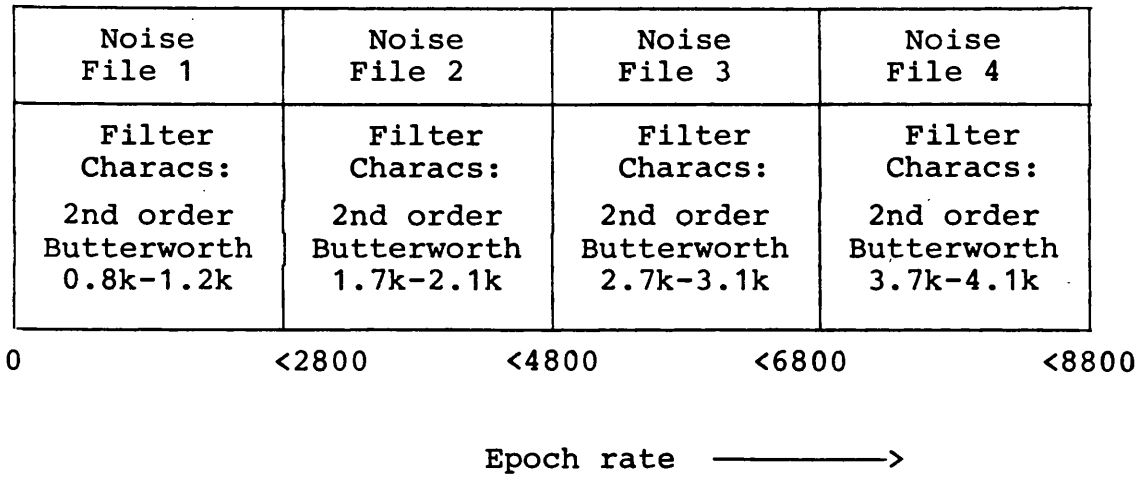


Figure 4.8 Subdivision of epoch rate variation range between 4 noise files

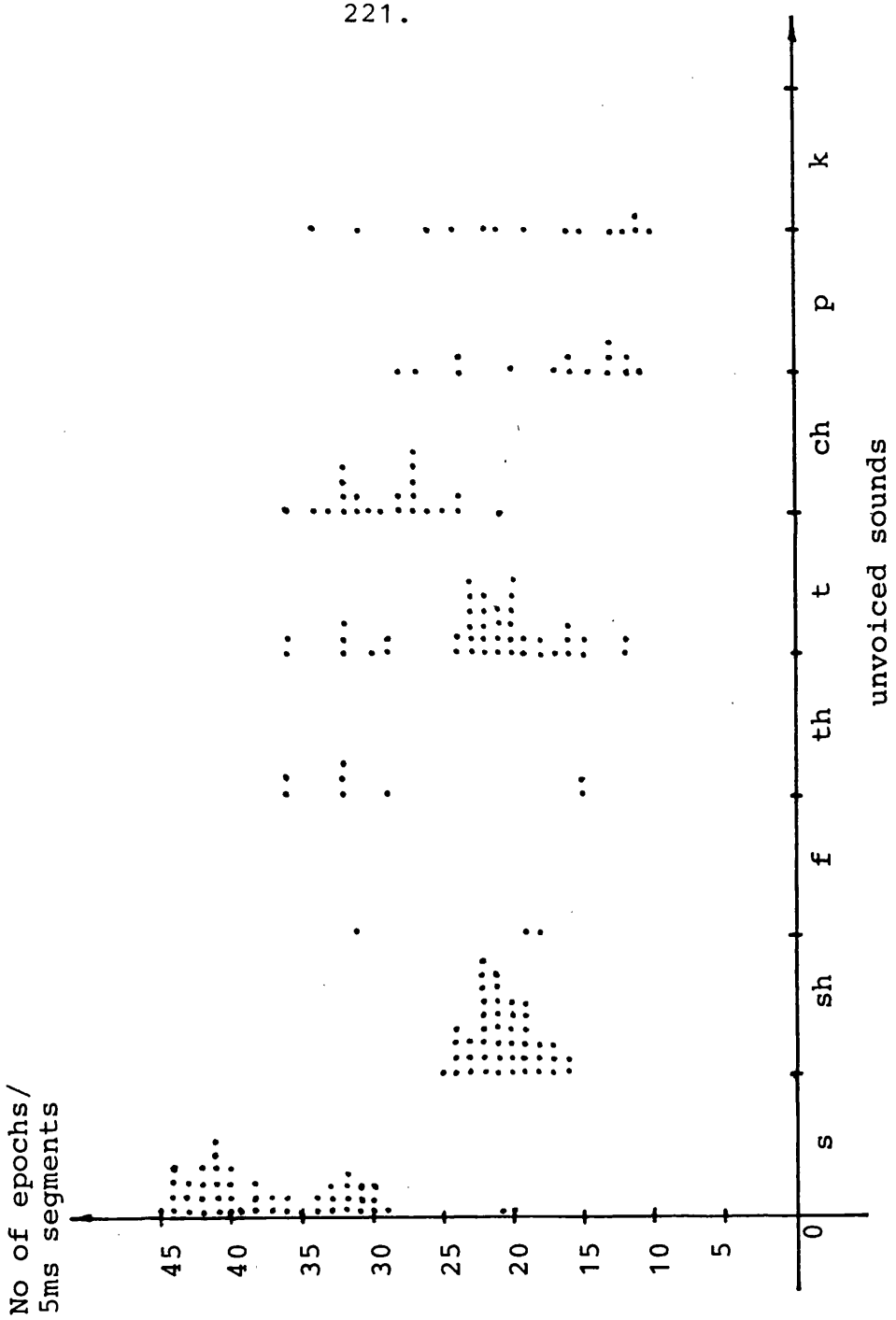
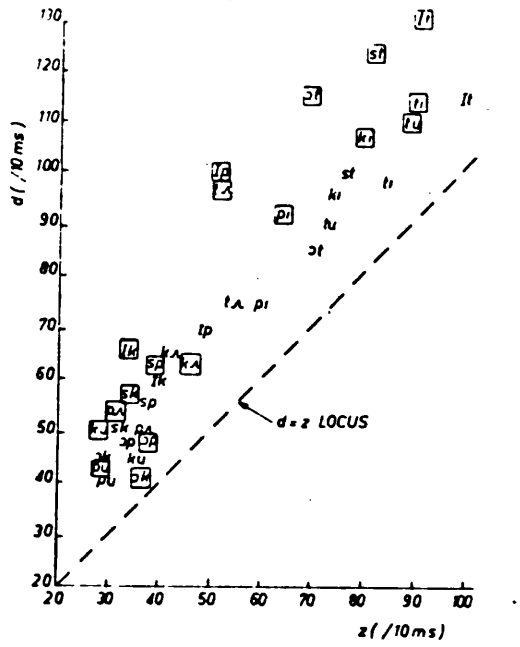
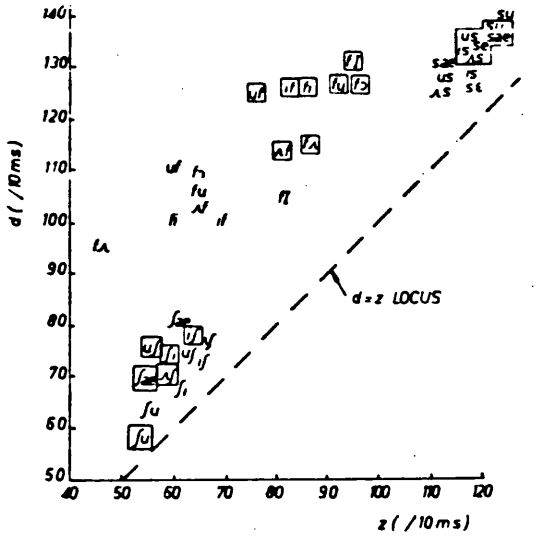


Figure 4.9 Measured epoch rates for the unvoiced sounds in the utterances of Appendix B



d versus z for unvoiced stop consonants measured in various contexts. All stops are in either word-initial or word-final position. Squares indicate d and z as calculated from spectral data.



d versus z for fricative consonants measured in various contexts. Fricatives are in either initial or final position in a word. Squares denote z and d as calculated from spectral data.

Figure 4.10 Epoch rate measurements for the different unvoiced sounds (from Ito and Donaldson)

d = No of epochs for differentiated speech

z = No of epochs for normal speech

Noise File 1	Noise File 2	Noise File 3	Noise File 4
Filter characs: 2nd order Butterworth low-pass with cut-off at 1.5kHz	Filter characs: 2nd order Chebyshev low-pass with 10dB ripple and cut-off at 3.3kHz	Filter characs: 2nd order Butterworth high-pass with cut-off at 3.0kHz	Filter characs: 3rd order Butterworth high-pass with cut-off at 4.0kHz

0 <2800 <4800 <6800 <8800

Epoch rate ——————>

Figure 4.11 Subdivision of epoch rate variation range between four spectrally shaped noise files

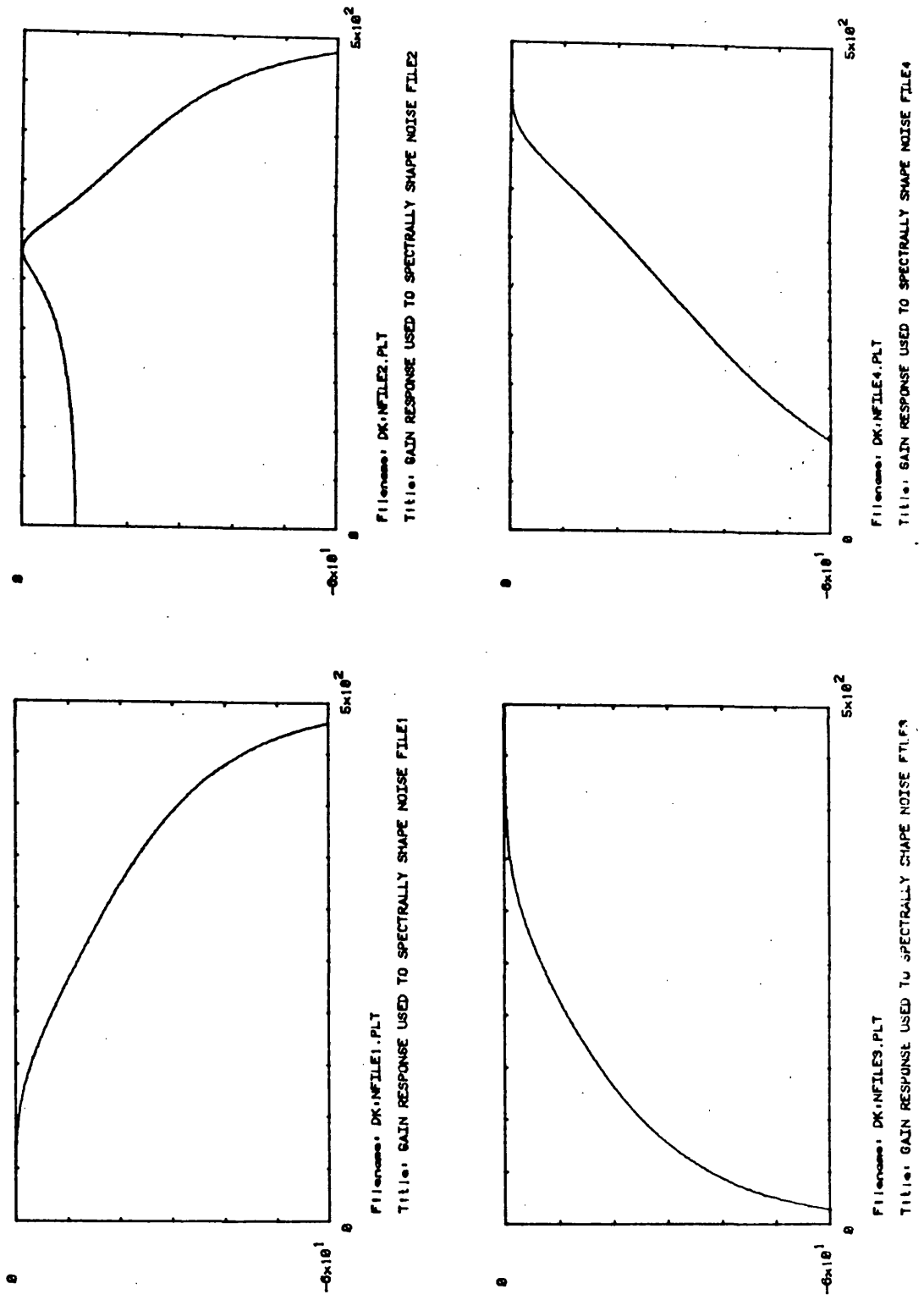


Figure 4.12 Gain response characteristics of the filters used to spectrally shape the noise files: vertical axes show gain in dBs; horizontal axes indicate one tenth of the frequency in hertz.

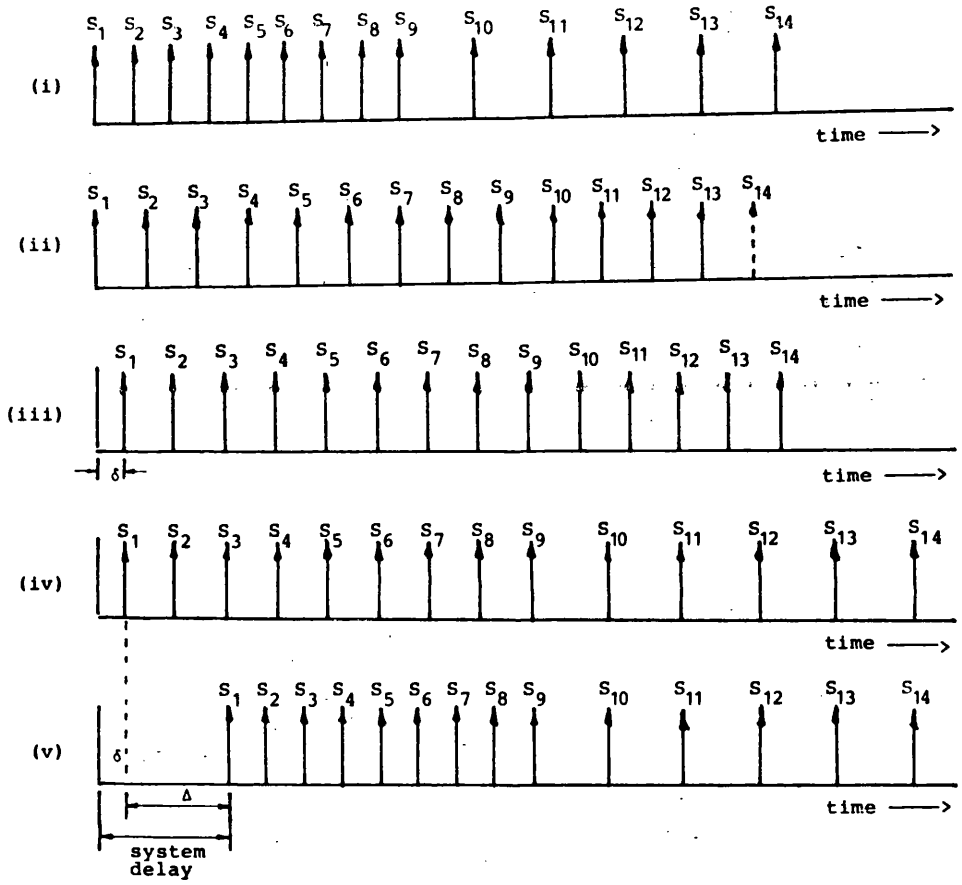


Figure 5.1 Illustration of output distortion due to insufficient decoding delay

(i) Input symbol rate
(ii) Constant transmission rate (transmission buffer underflow, S_{14})
(iii) Constant transmission rate (buffer underflow avoided by delay, δ)
(iv) System output (distortion due to insufficient delays)
(v) System output (distortion avoided by including delay Δ)

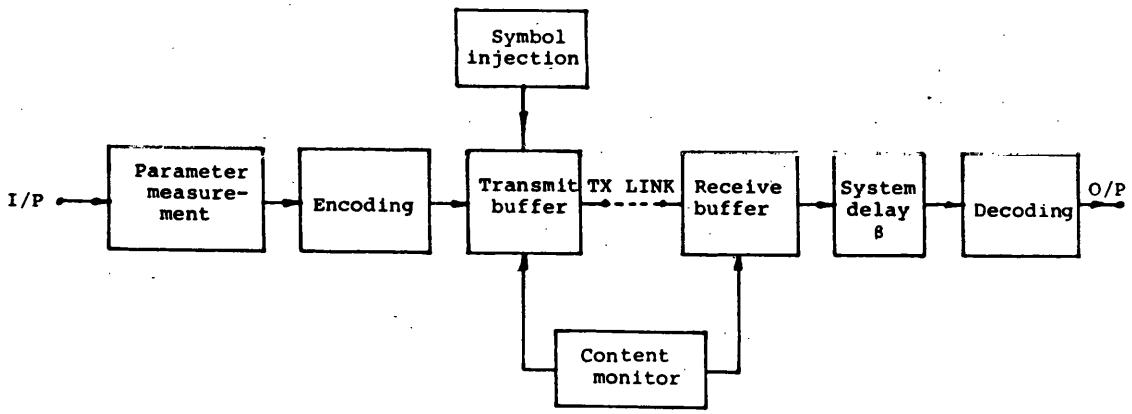


Figure 5.2 System for measuring the delay and buffer size requirements

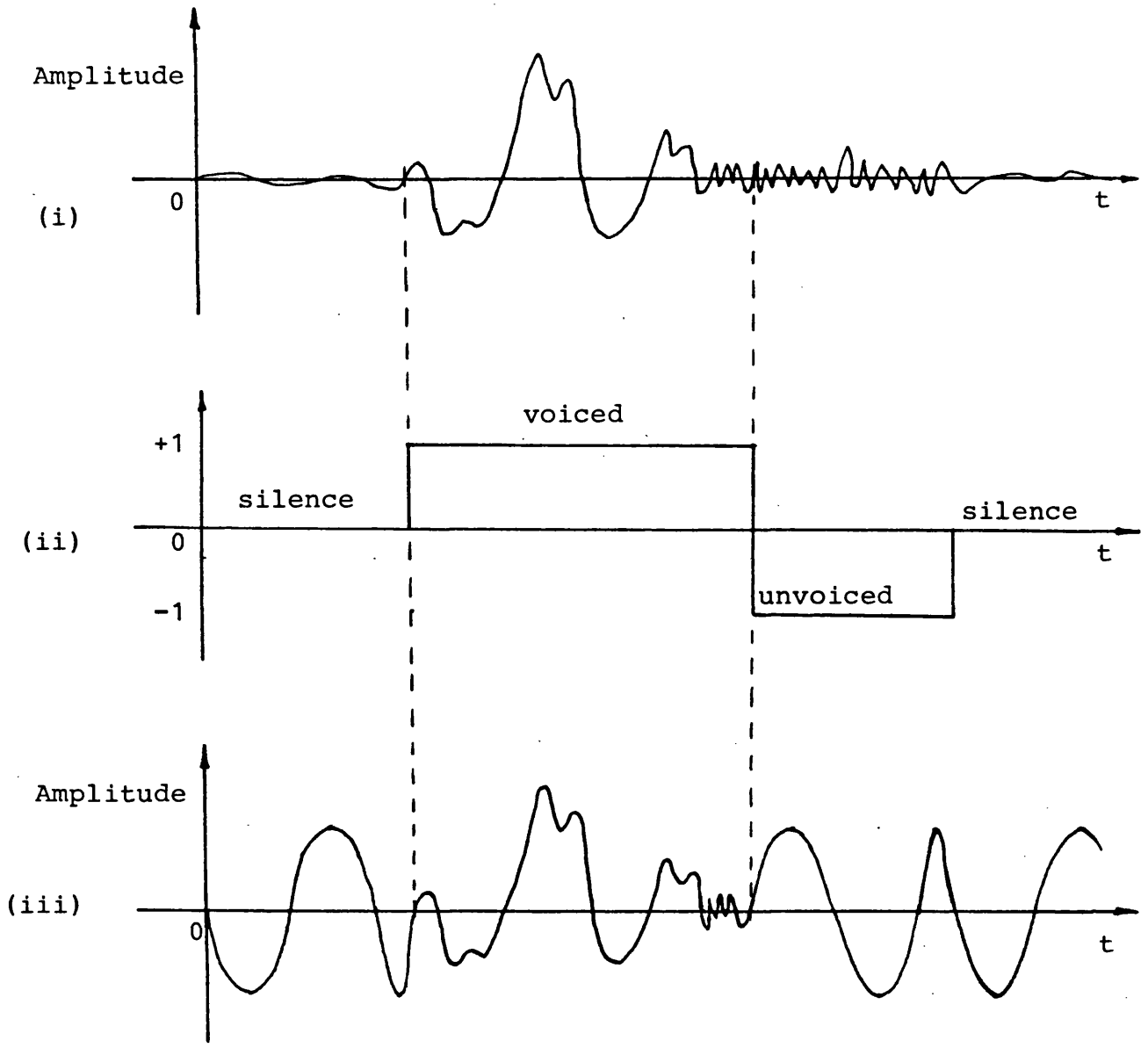


Figure 5.3 Making inputs compatible for system used to measure delay and buffer size

(i) TES input to measuring system
(ii) Classification of the waveform
(iii) Hybrid-TES input to measuring system

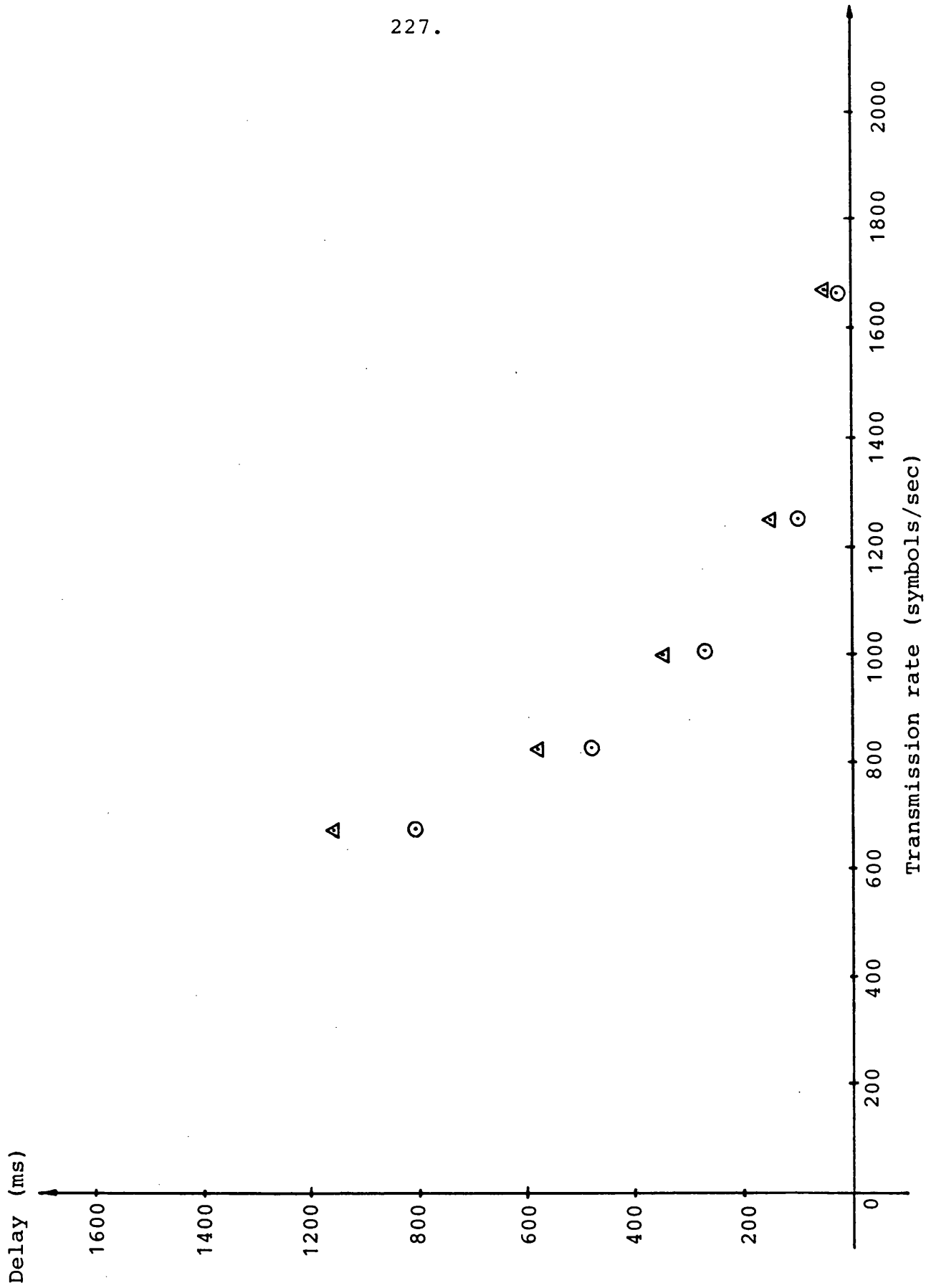


Figure 5.4 Minimum delay requirements at different tranmission rates for the speech file FEM1.SPH

△ TES

⊙ Hybrid-TES

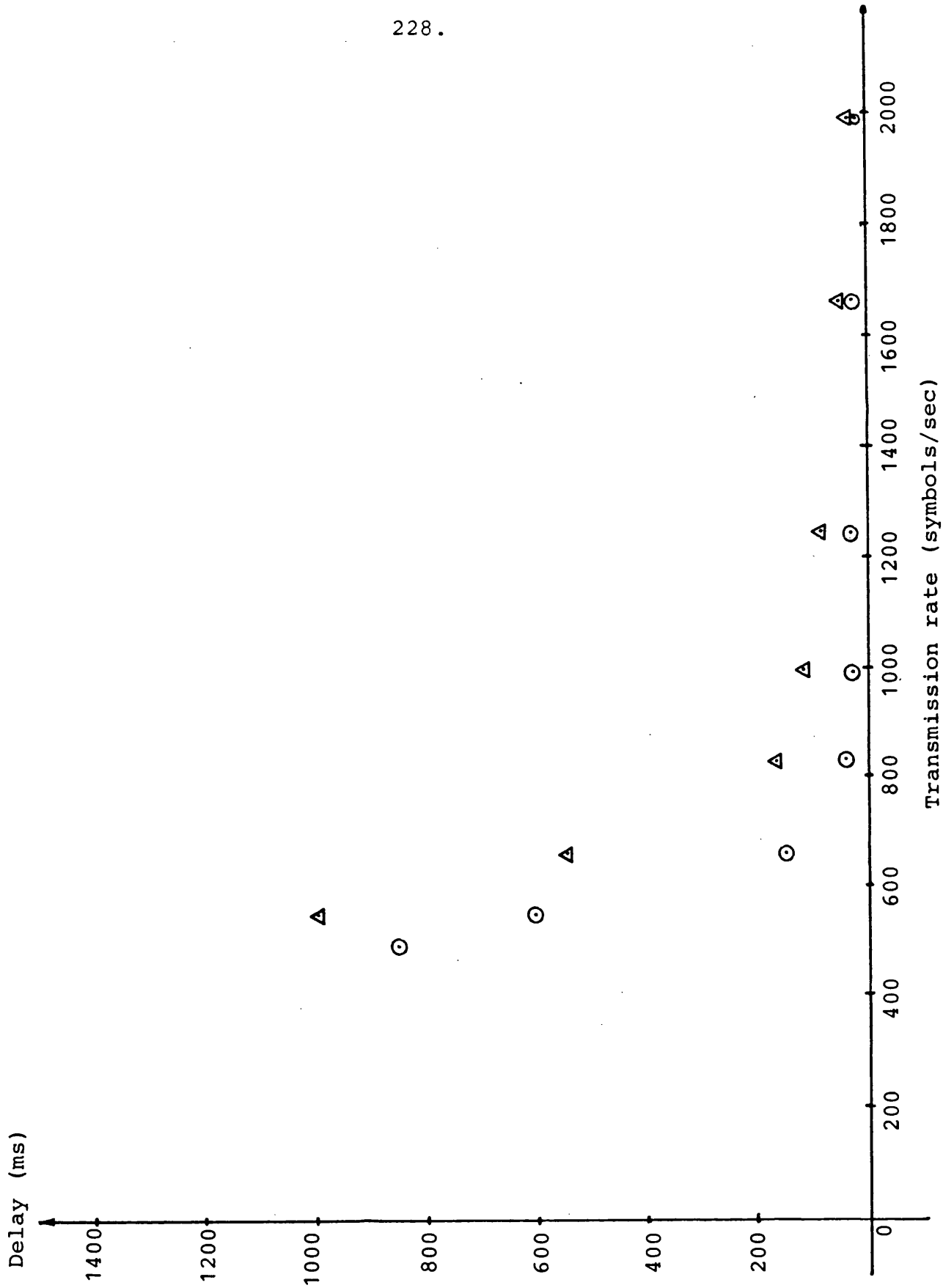


Figure 5.5 Minimum delay requirements at different transmission rates for the speech file APPLE8.SPH

△ TES

⊙ Hybrid-TES

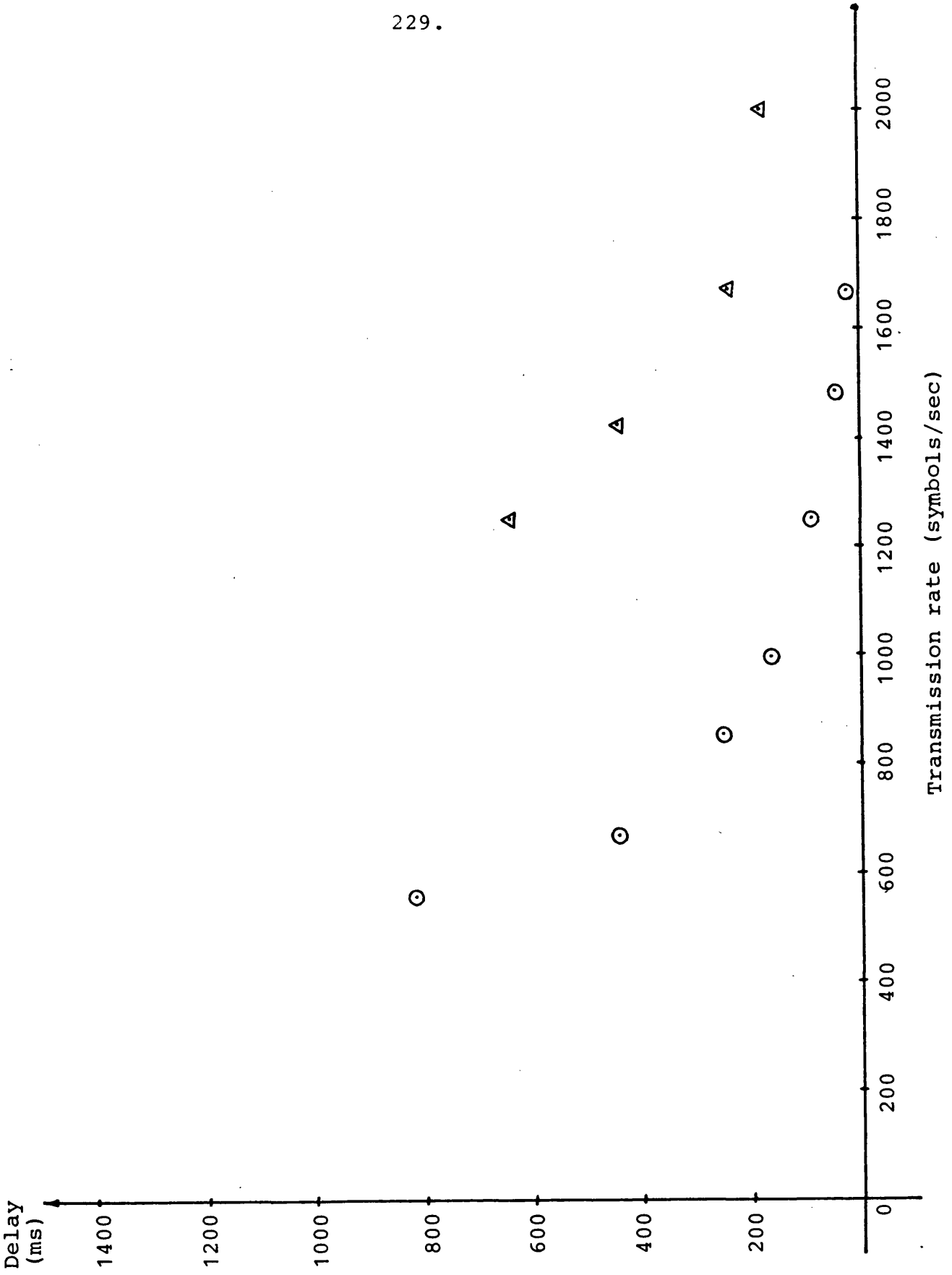


Figure 5.6 Minimum delay requirements at different transmission rates for the speech file CBONLY.SPH

△ TES

⊙ Hybrid-TES

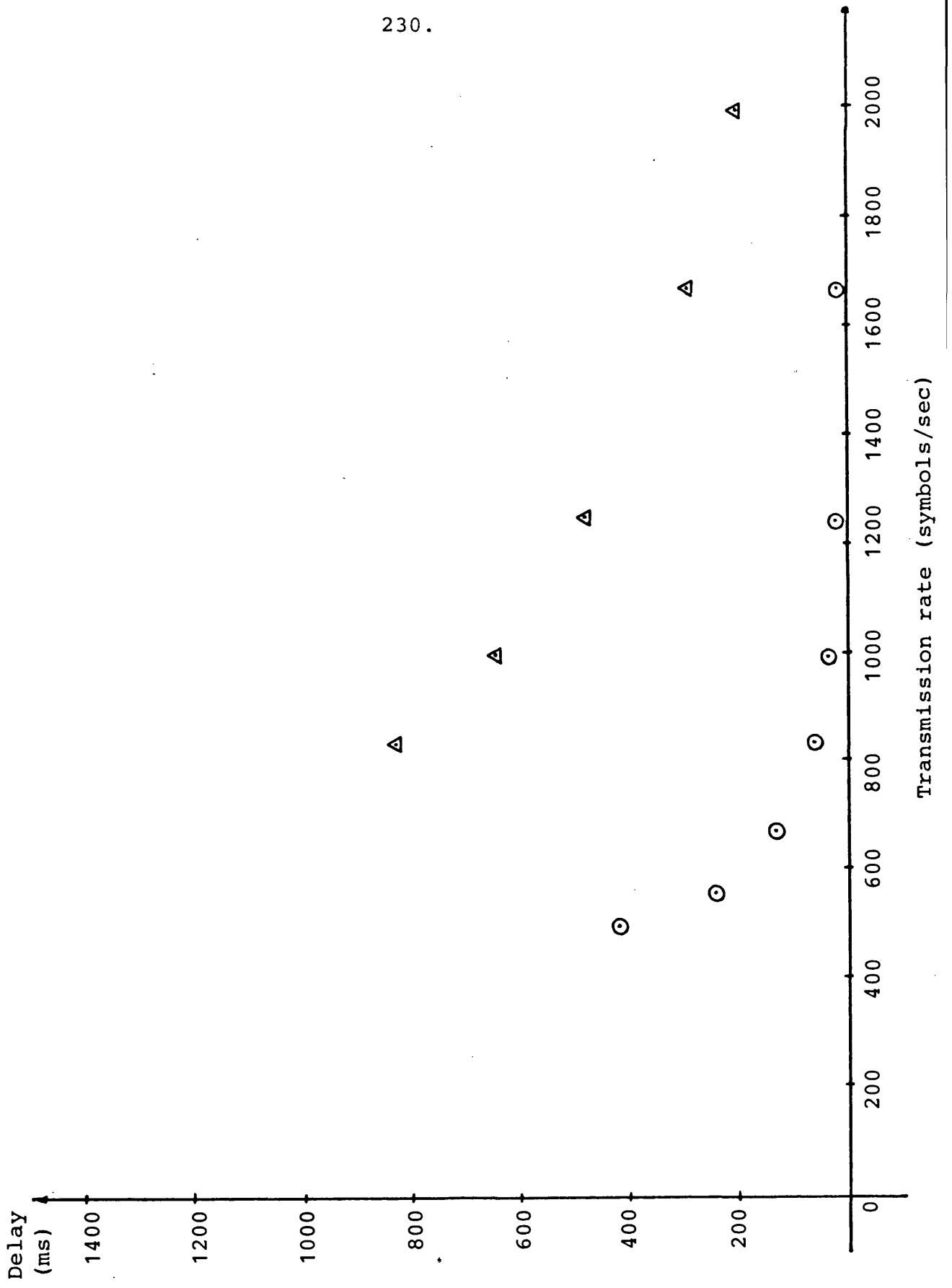


Figure 5.7 Minimum delay requirements at different transmission rates for the speech file BIRD.SPH

△ TES

⊙ Hybrid-TES

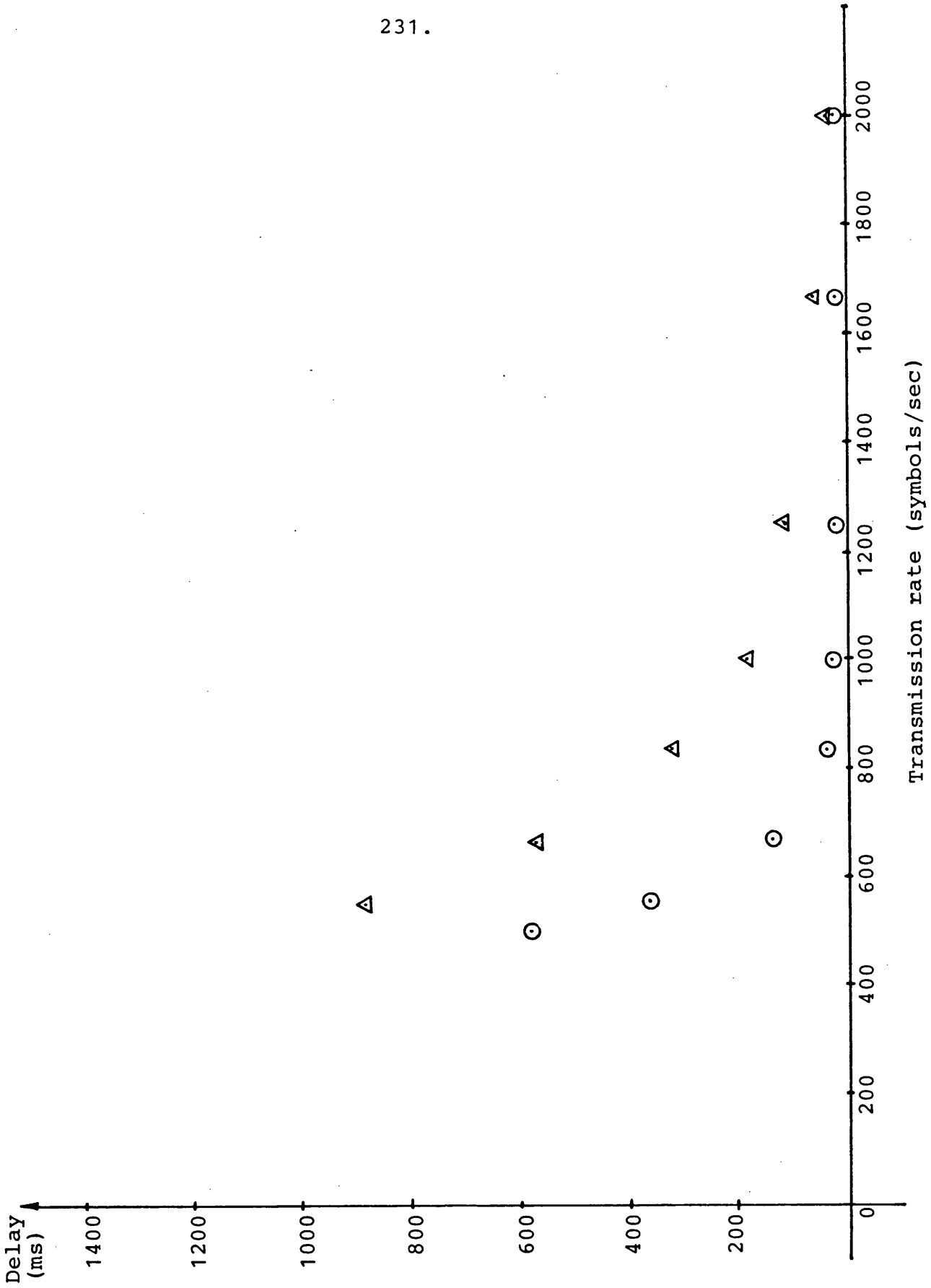


Figure 5.8 Minimum delay requirements at different transmission rates for the speech file APPLE7.SPH

△ TES

⊙ Hybrid-TES

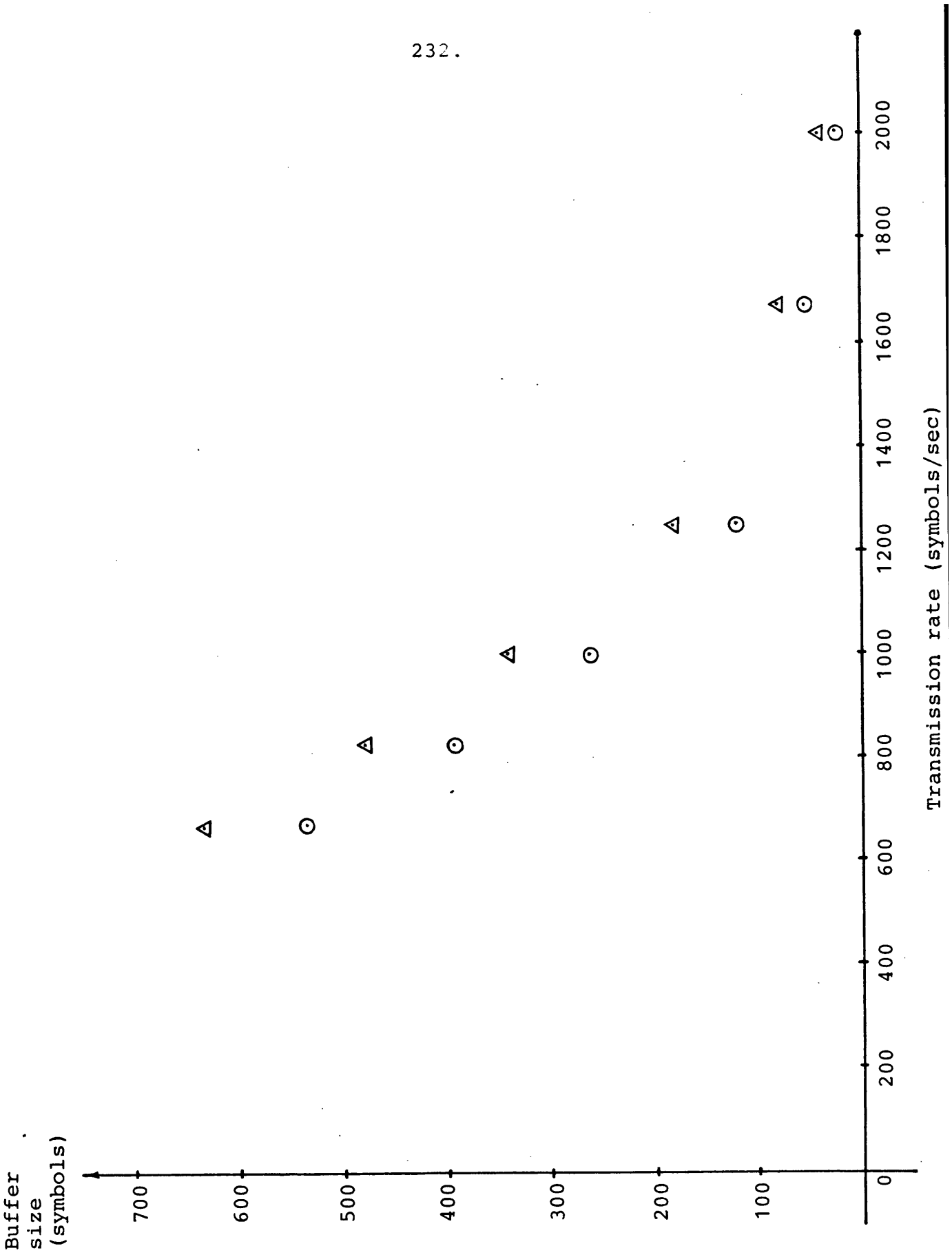


Figure 5.9 Minimum buffer size requirements for no repeats of the speech file FEM1.SPH at different transmission rates

△ TES

⊙ Hybrid-TES

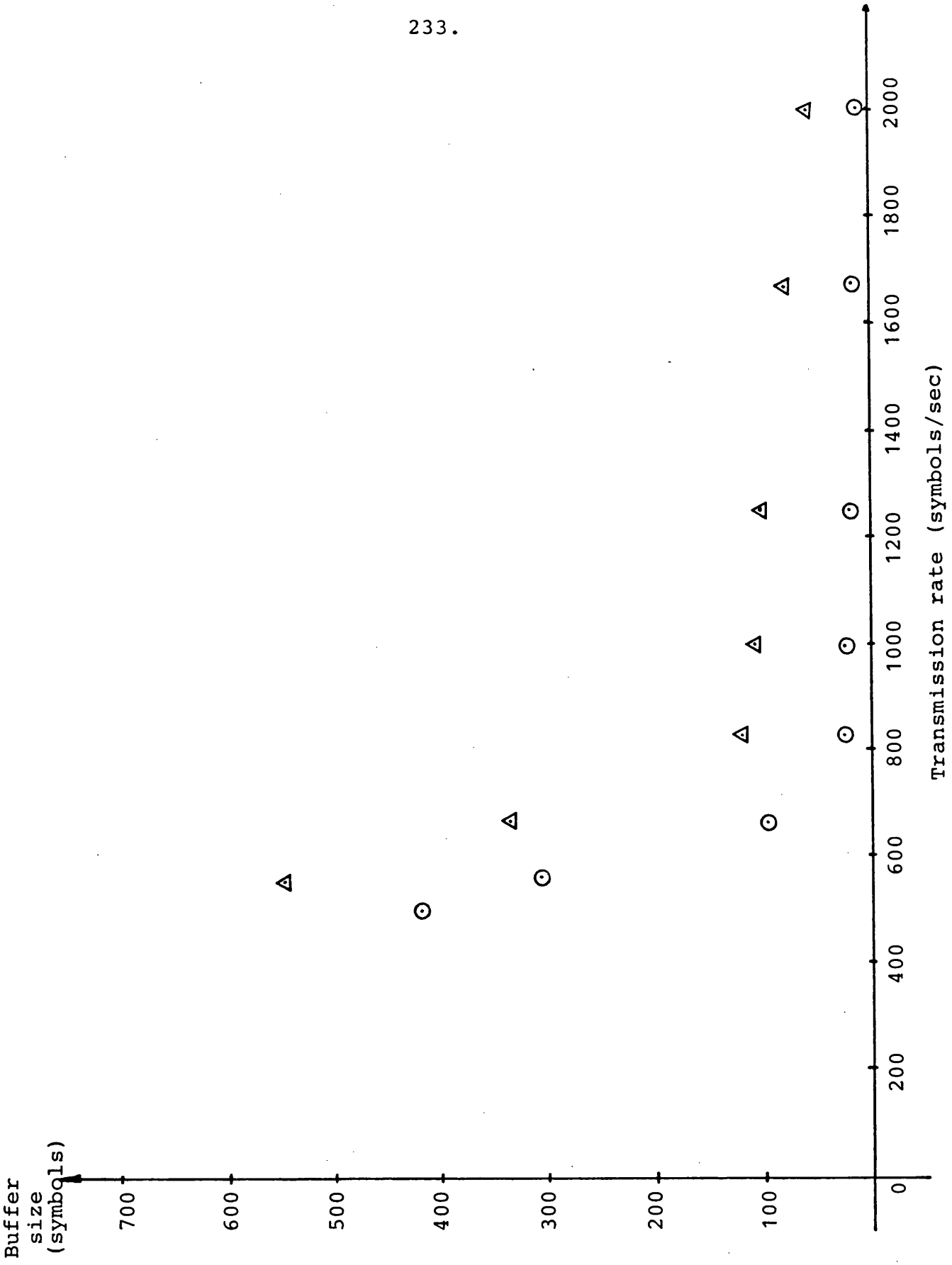


Figure 5.10 Minimum buffer size requirements for no repeats of the speech file APPLE8.SPH at different transmission rates

△ TES

⊙ Hybrid-TES

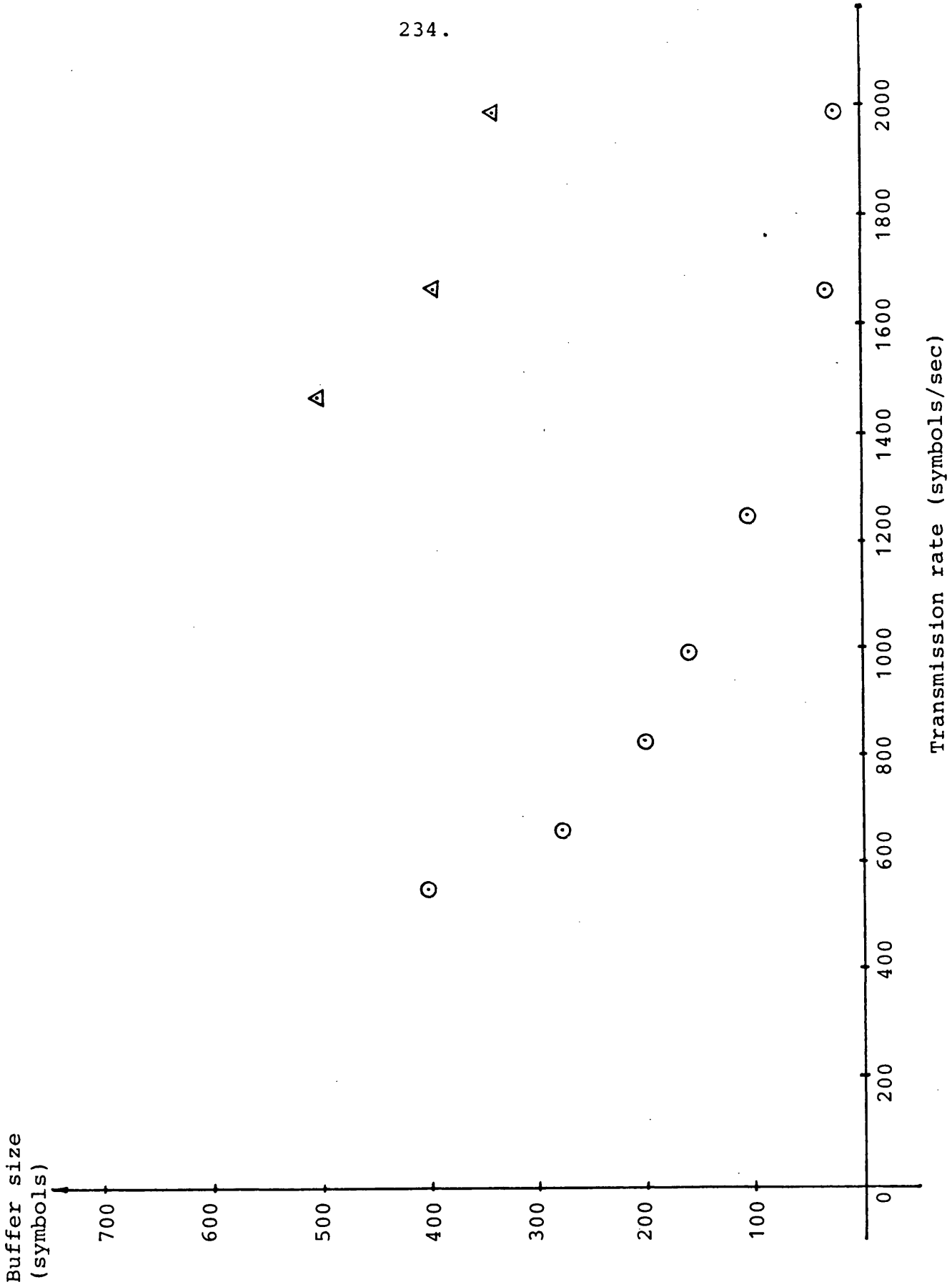


Figure 5.11 Minimum buffer size requirements for no repeats of the speech file CBONLY.SPH at different transmission rates

△ TES

⊙ Hybrid-TES

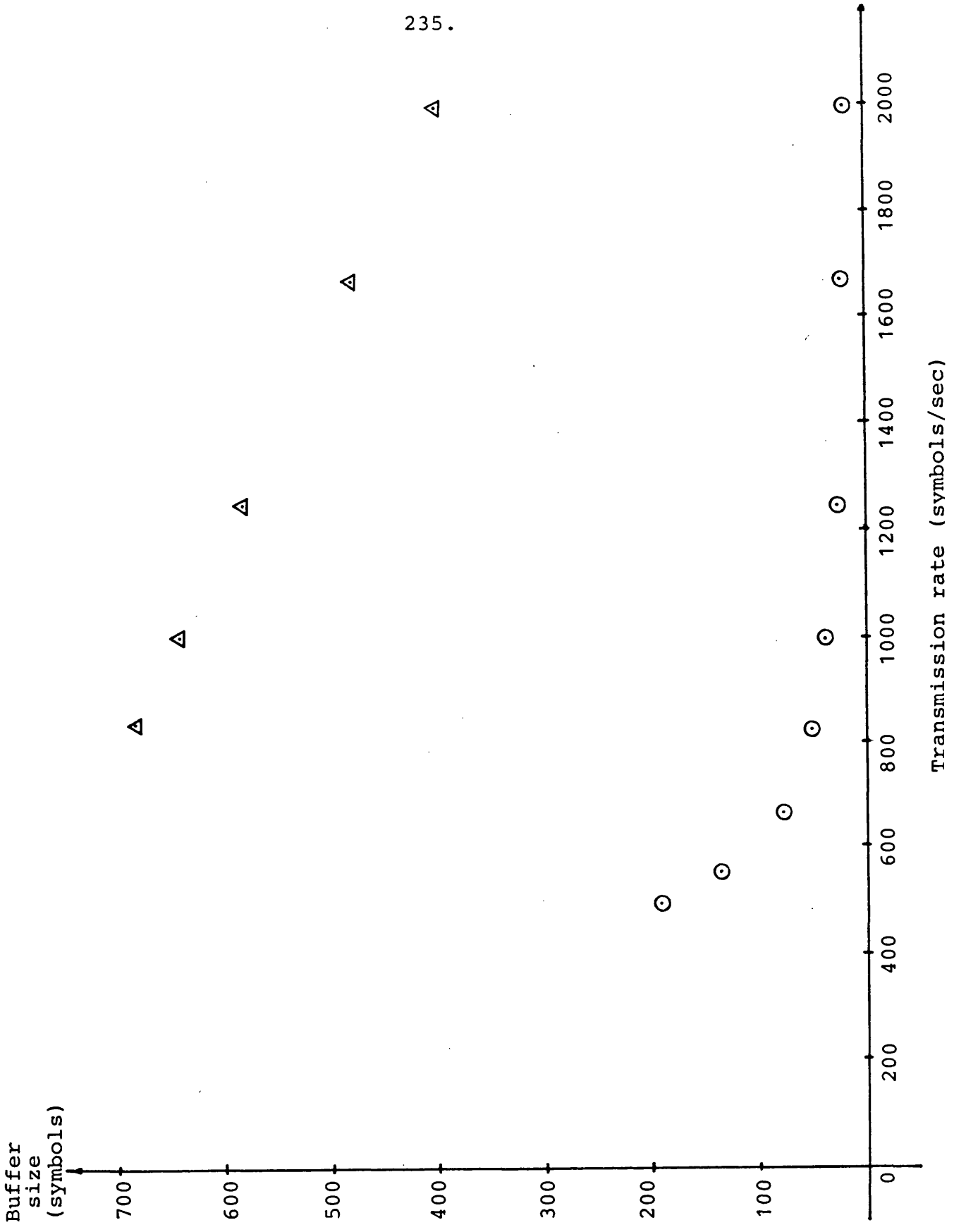


Figure 5.12 Minimum buffer size requirements for no repeats of the speech file BIRD.SPH at different transmission rates

△ TES

⊙ Hybrid-TES

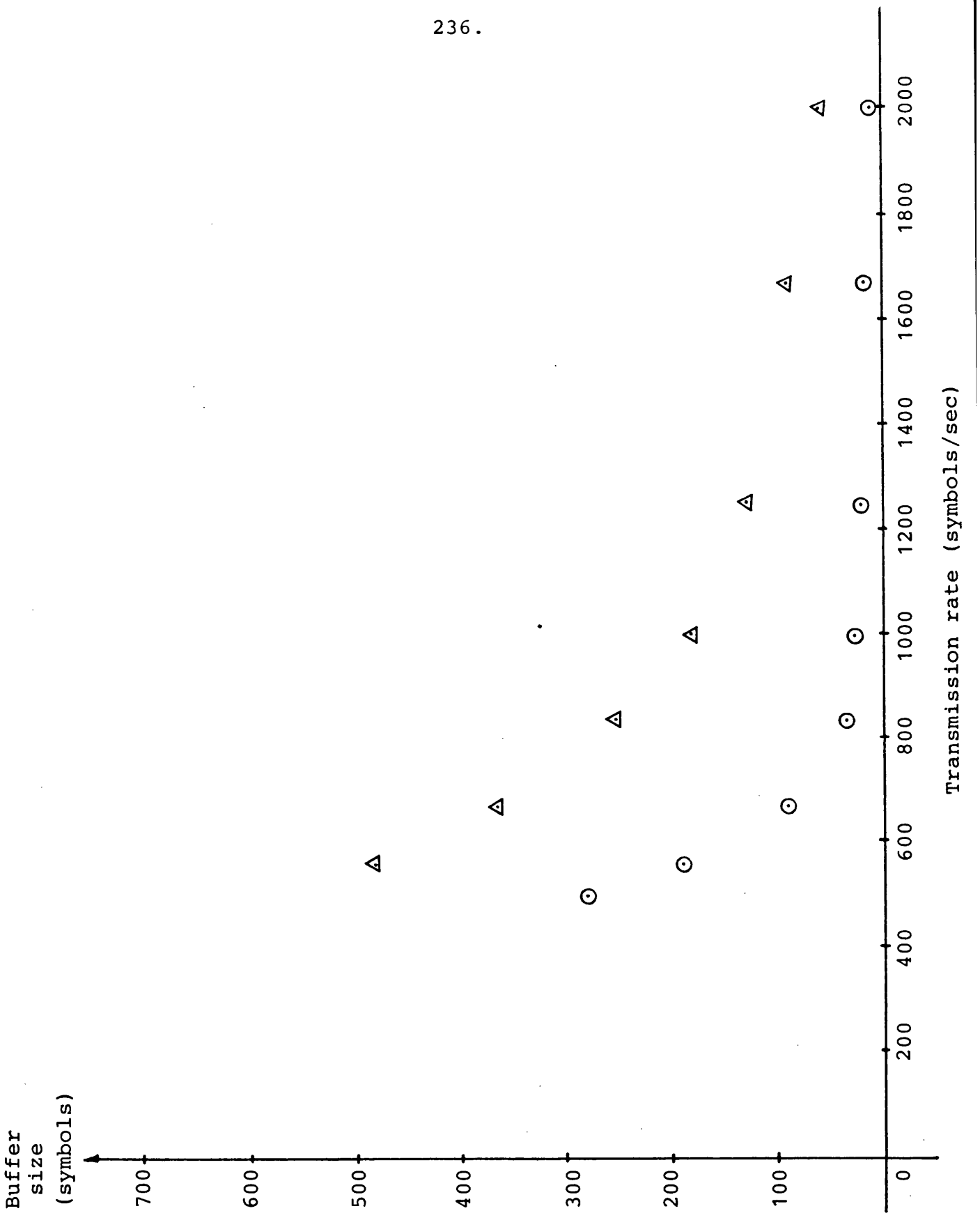


Figure 5.13 Minimum buffer size requirements for no repeats of the speech file APPLE7.SPH at different transmission rates

△ TES

⊙ Hybrid-TES

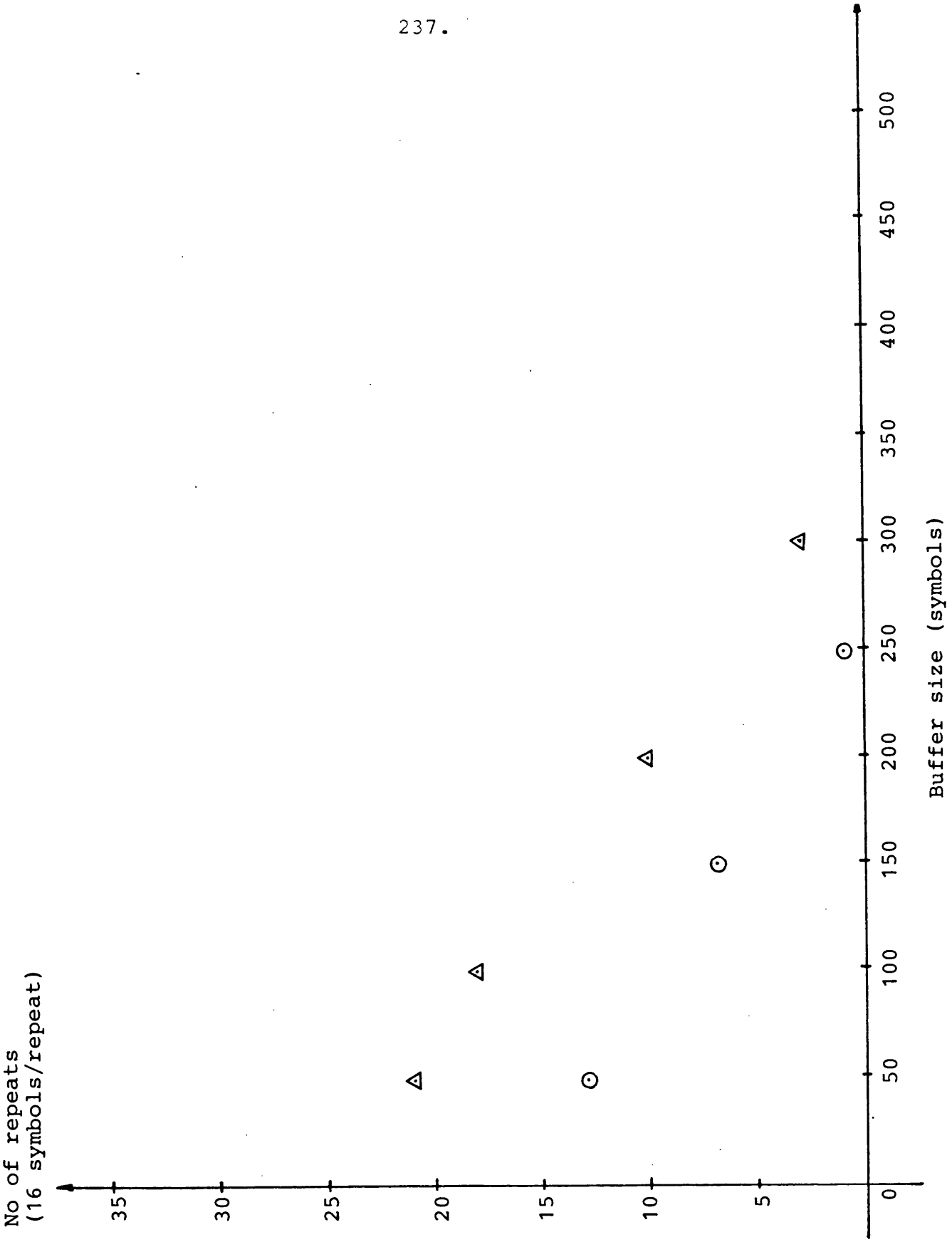


Figure 5.14 Number of repeats necessary with varying buffer size for the speech file FEM1.SPH. Transmission rate: 1000 symbols/sec

△ TES

⊙ Hybrid-TES

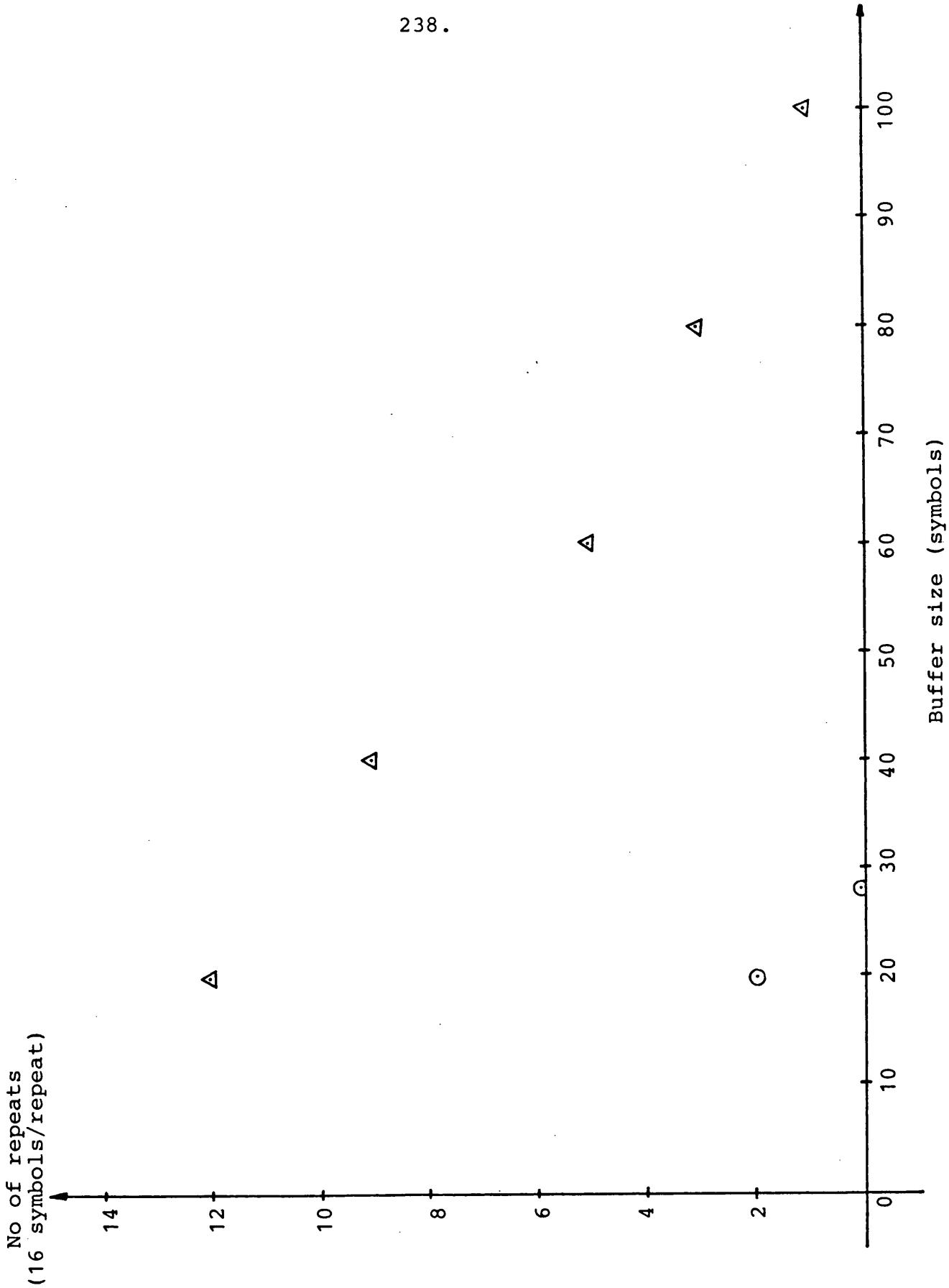


Figure 5.15 Number of repeats necessary with varying buffer size for the speech file APPLE8.SPH. Transmission rate: 1000 symbols/sec.

△ TES

⊙ Hybrid-TES

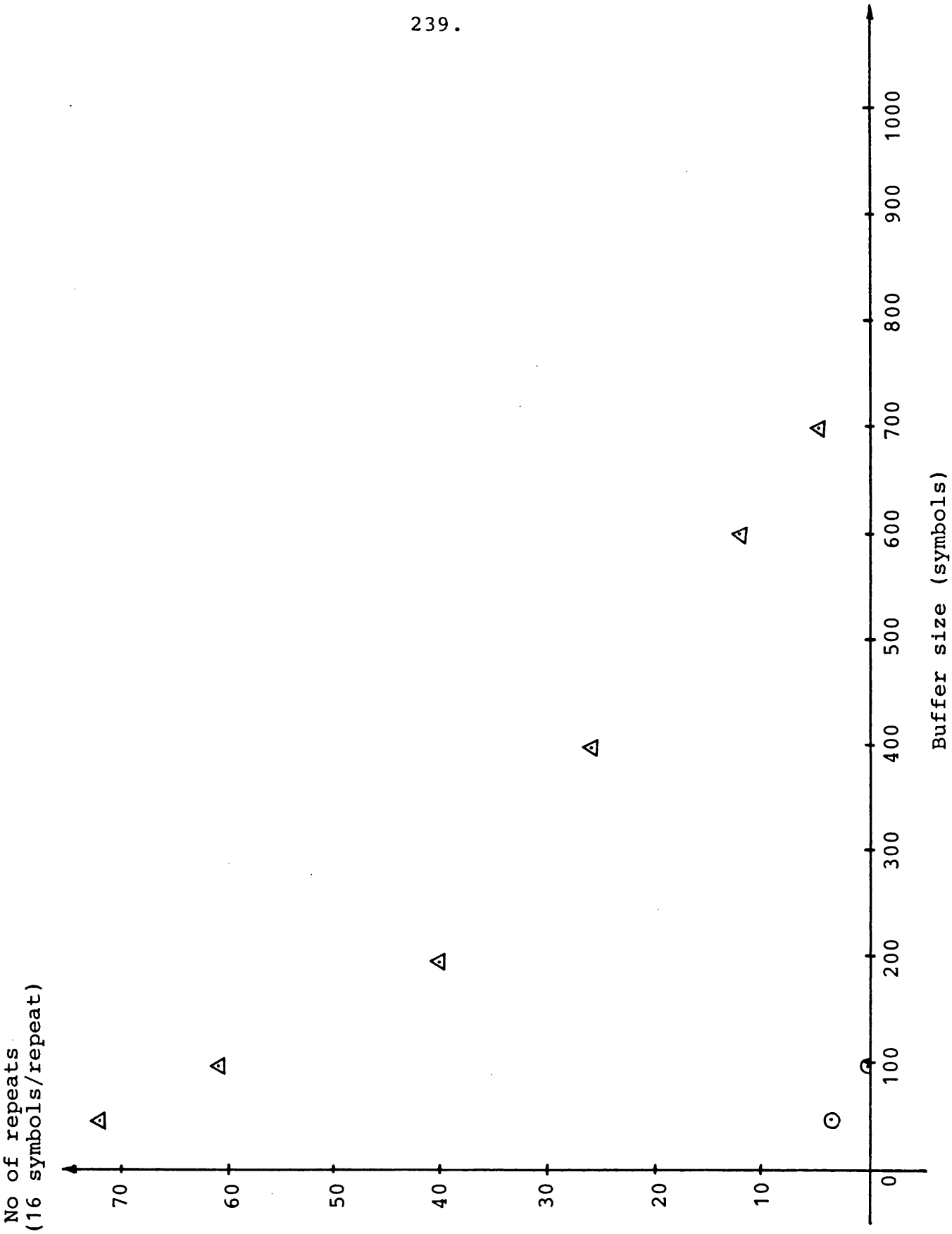


Figure 5.16 Number of repeats necessary with varying buffer size for the speech file CBONLY.SPH. Transmission rate: 1250 symbols/sec.

△ TES

⊙ Hybrid-TES

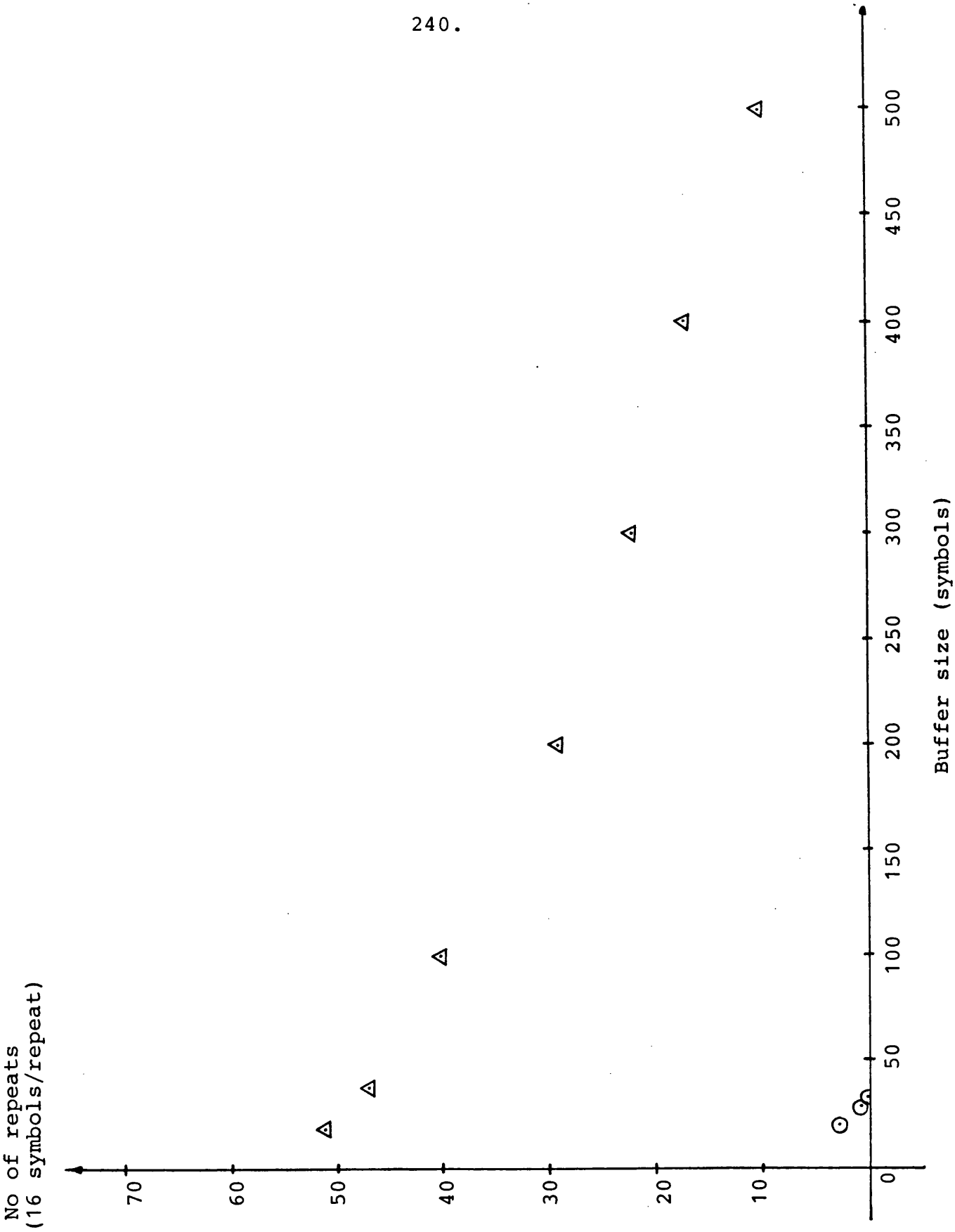


Figure 5.17 Number of repeats necessary with varying buffer size for the speech file BIRD.SPH. Transmission rate: 1000 symbols/sec.

△ TES

⊙ Hybrid-TES

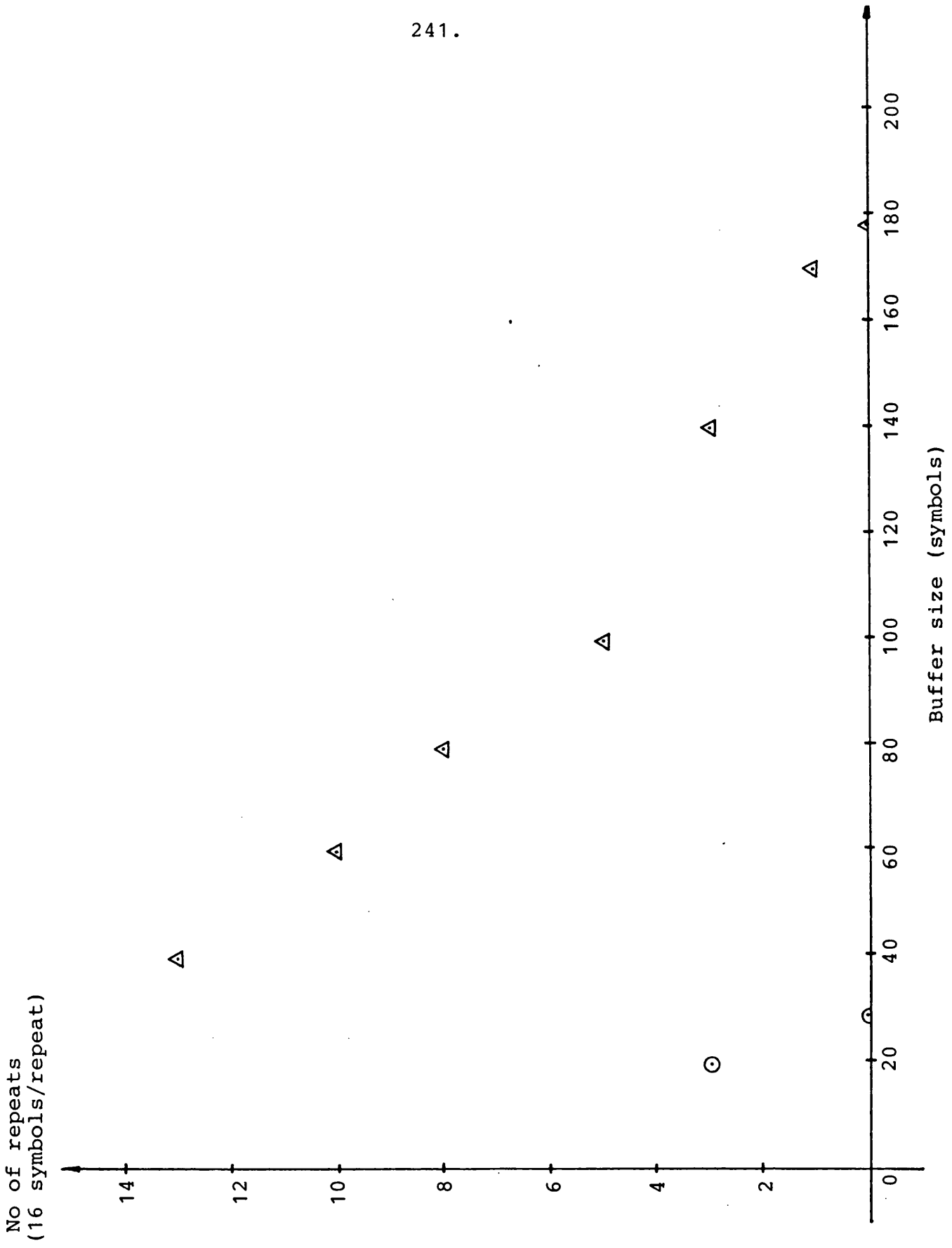


Figure 5.18 Number of repeats necessary with varying buffer size for the speech file APPLE7.SPH. Transmission rate: 1000 symbols/sec

△ TES

⊙ Hybrid-TES

Comparison	Actual file designation, as stored in computer memory				
(1) $SSHT_A$-V-RST_A	DL0:APPLE8.RE2	&	DL1:	-	.BFF
(2) RST_B-V-BHT_B	DL1:APPLE7.BFF	&	DL0:	-	.RE1
(3) $SSHT_G$-V-BHT_G	DL1:CB33K4.RE2	&	DL1:	-	.RE1
(4) BHT_D-V-RST_D	DL0:BIRD.RE1	&	DL0:	-	.BF1
(5) $SSHT_G$-V-BHT_G	DL1:B33K4.RE2	&	DL1:	-	.RE1
(6) RST_C-V-$SSHT_C$	DL0:CBONLY.BF1	&	DL0:	-	.RE2
(7) PCM -V-NOISY	DL0:APPLE8.SPH	&	DL1:	A8P30	.SPN
(8) $SSHT_F$-V-BHT_F	DL0:FEM1.RE2	&	DL0:	-	.RE1
(9) RST_C-V-$SSHT_C$	DL0:BIRD.BF1	&	DL0:	-	.RE2
(10) $SSHT_C$-V-RST_C	DL0:CBONLY.RE2	&	DL0:	-	.BF1
(11) RST_B-V-BHT_B	DL1:BIRD.BFF	&	DL0:	-	.RE1
(12) BHT_F-V-$SSHT_F$	DL0:APPLE8.RE1	&	DL0:	-	.RE2
(13) RST_D-V-BHT_D	DL0:CBONLY.BF1	&	DL0:	-	.RE1
(14) NOISY-V-PCM	DL1:A8P30.SPN	&	DL0:	APPLE8	.SPH
(15) BHT_B-V-RST_B	DL0:APPLE7.RE1	&	DL0	-	.BFF
(16) $SSHT_A$-V-RST_A	DL0:CBONLY.RE2	&	DL1	-	.BFF
(17) BHT_G-V-$SSHT_G$	DL1:B33K4.RE1	&	DL1:	-	.RE2
(18) BHT -V-RST_B	DL0:BIRD.RE1	&	DL1:	-	.BFF
(19) RST_B-V-BHT_B	DL1:FEM1.BFF	&	DL0:	-	.RE1
(20) BHT_F-V-$SSHT_F$	DL0:CBONLY.RE1	&	DL0:	-	.RE2
(21) PCM -V-NOISY	DL0:APPLE8.SPH	&	DL1:	A8P30	.SPN
(22) RST_D-V-BHT_D	DL0:BIRD.BF1	&	DL0:	-	.RE1
(23) BHT_B-V-RST_B	DL0.CBONLY.RE1	&	DL1:	-	.BFF
(24) $SSHT_A$-V-RST_A	DL0:FEM1.RE2	&	DL1:	-	.BFF
(25) RST_A-V-$SSHT_A$	DL1:APPLE8.BFF	&	DL0:	-	.RE2
(26) $SSHT_F$-V-BHT_F	DL0:APPLE7.RE2	&	DL0:	-	.RE1
(27) $SSHT_E$-V-RST_E	DL0.CBONLY.RE2	&	DL0:	-	.BF2
(28) NOISY-V-PCM	DL1:A8P30.SPN	&	DL0:	APPLE8	.SPH
(29) $SSHT_C$-V-RST_C	DL0:BIRD.RE2	&	DL0:	-	.BF1
(30) BHT_B-V-RST_B	DL0.FEM1.RE1	&	DL1:	-	.BFF

Comparison	Actual file designation, as stored in computer memory				
(31) $SSHT_F$-V-BHT_F	DL0:CBONLY.RE2	&	DL0:	-	.RE1
(32) BHT_B-V-RST_B	DL0:APPLE8.RE1	&	DL1:	-	.BFF
(33) RST_A-V-$SSHT_A$	DL1:CBONLY.BFF	&	DL0:	-	.RE2
(34) BHT_F-V-$SSHT_F$	DL0:FEM1.RE1	&	DL0:	-	.RE2
(35) PCM -V-NOISY	DL0:APPLE8.SPH	&	DL1:	A8P30	.SPN
(36) BHT_G-V-$SSHT_G$	DL1:CB33K4.RE1	&	DL1:	-	.RE2
(37) BHT_F-V-$SSHT_F$	DL0:BIRD.RE1	&	DL0:	-	.RE2
(38) $SSHT_F$-V-BHT_F	DL0:APPLE8.RE2	&	DL0:	-	.RE1
(39) RST_B-V-BHT_B	DL1:CBONLY.BFF	&	DL0:	-	.RE1
(40) $SSHT_F$-V-BHT_F	DL0:BIRD.RE2	&	DL0:	-	.RE1
(41) RST_A-V-$SSHT_A$	DL1:APPLE7.BFF	&	DL0:	-	.RE2
(42) NOISY-V-PCM	DL1:A8P30.SPN	&	DL0:	APPLE8	.SPH
(43) RST_B-V-BHT_B	DL1:APPLE8.BFF	&	DL0:	-	.RE1
(44) RST_E-V-$SSHT_E$	DL0:CBONLY.BF2	&	DL0:	-	.RE2
(45) BHT_F-V-$SSHT_F$	DL0:APPLE7.RE1	&	DL0:	-	.RE2
(46) $SSHT_A$-V-RST_A	DL0:BIRD.RE2	&	DL1:	-	.BFF
(47) BHT_D-V-RST_D	DL0:CBONLY.RE1	&	DL0:	-	.BF1
(48) RST_A-V-$SSHT_A$	DL1:FEM1.BFF	&	DL0:	-	.RE2
(49) RST_A-V-$SSHT_A$	DL1:BIRD.BFF	&	DL0:	-	.RE2
(50) $SSHT_A$-V-RST_A	DL0:APPLE7.RE2	&	DL1:	-	.BFF

Figure 5.19 Order of presentation of comparisons to listeners

AS SUBJECTIVE LISTENING TESTS 8.12.81

You will hear a sequence of pairs of sentences and you should state which one (if any) of the pair you find more acceptable. So as to familiarise you with the procedure and enable you to set the volume to a desirable level, a practice pair is provided. Indicate your preference by placing a tick in the appropriate column adjacent to the pair you are listening to, e.g.

Pair No	Code Word	Sentence 1	Sentence 2
10	Apple	✓	

indicates preference of sentence 1.

If there is no preference please tick both columns. The code word gives an indication of the content of the sentence pair. This is to assist you in keeping track and marking the appropriate row.

Practice Pair	SWIMMING		
1	APPLE		
2	APPLE		
3	CHARLES		
4	BIRD		
5	BIRD		
6	CHARLES		
7	APPLE		
8	SWIMMING		
9	BIRD		
10	CHARLES		
11	BIRD		
12	APPLE		
13	CHARLES		
14	APPLE		
15	APPLE		
16	CHARLES		
17	BIRD		
18	BIRD		
19	SWIMMING		
20	CHARLES		
21	APPLE		
22	BIRD		
23	CHARLES		
24	SWIMMING		
25	APPLE		
26	APPLE		
27	CHARLES		
28	APPLE		
29	BIRD		
30	SWIMMING		
31	CHARLES		
32	APPLE		
33	CHARLES		
34	SWIMMING		
35	APPLE		
36	CHARLES		
37	BIRD		
38	APPLE		
39	CHARLES		
40	BIRD		
41	APPLE		
42	APPLE		
43	APPLE		
44	CHARLES		
45	APPLE		
46	BIRD		
47	CHARLES		
48	SWIMMING		
49	BIRD		
50	APPLE		

Figure 5.20 The instruction sheet as presented to the listeners during the subjective listening tests

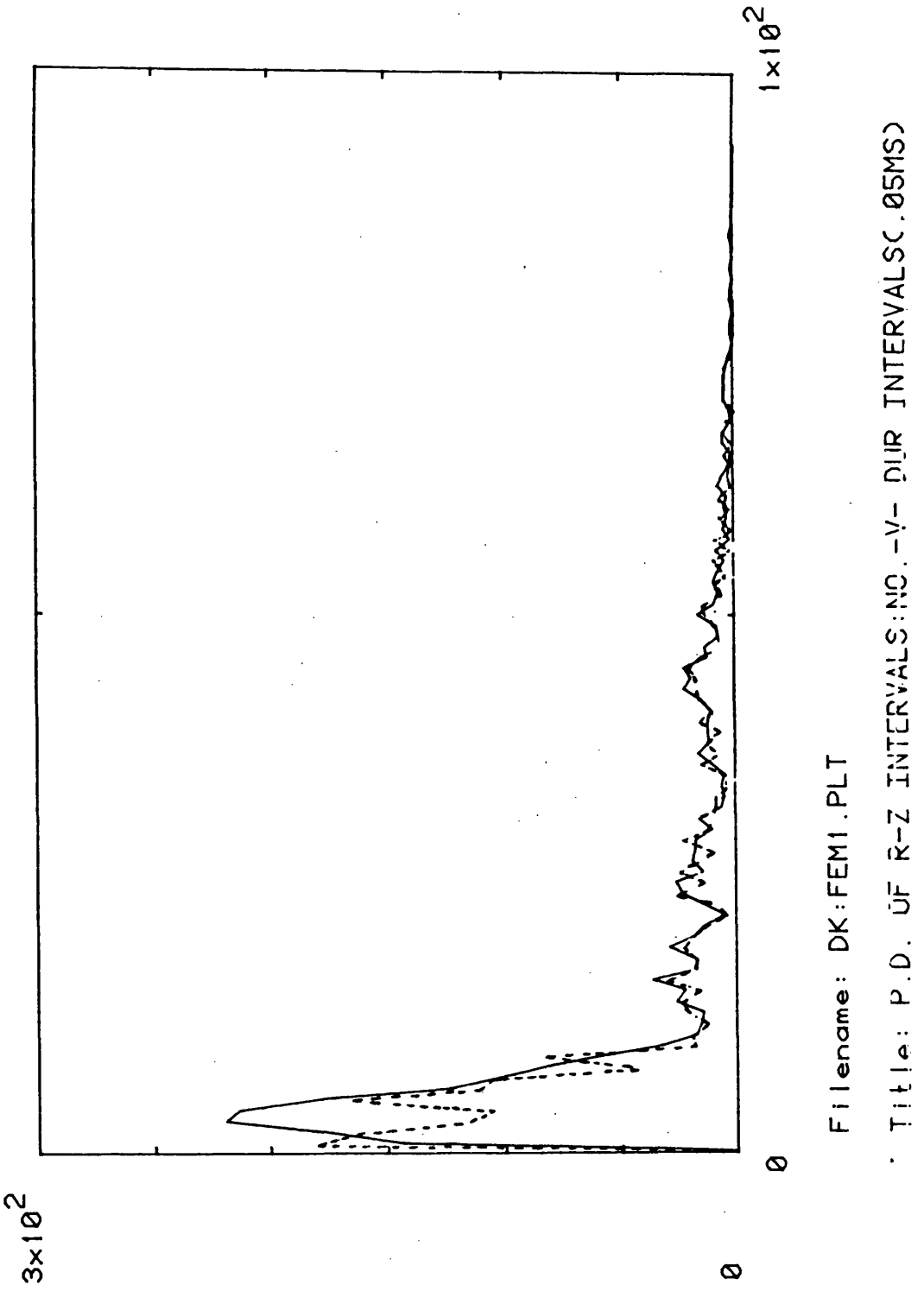


Figure 5.21

Comparison of the probability distributions for speech file FEM1.SPI!

_____ the original real-zero intervals

------- symbols stored in the transmitter buffer

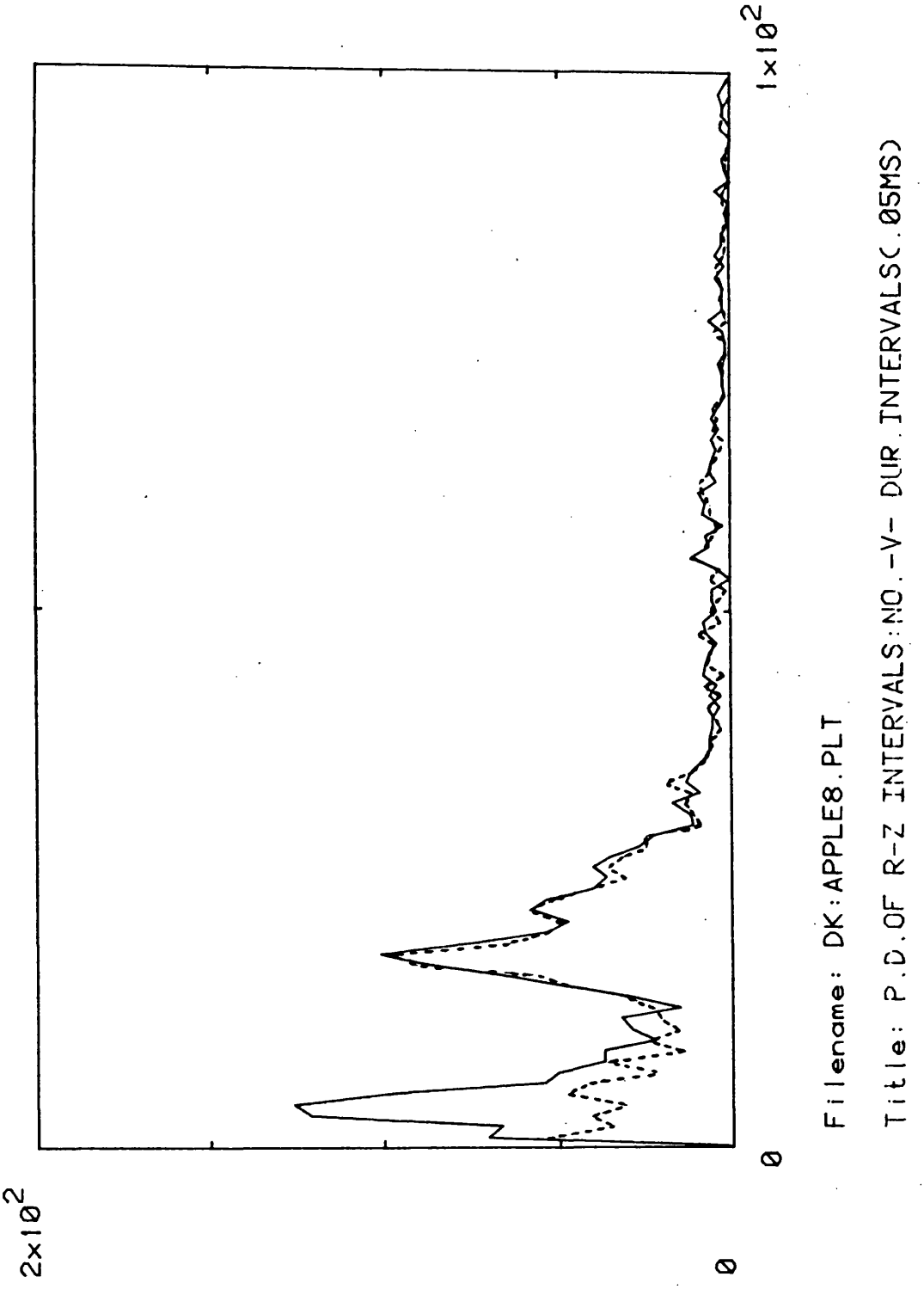


Figure 5.22

Comparison of the probability distributions for speech file APPLE8.SPH

——— the original real-zero intervals

------- symbols stored in the transmitter buffer

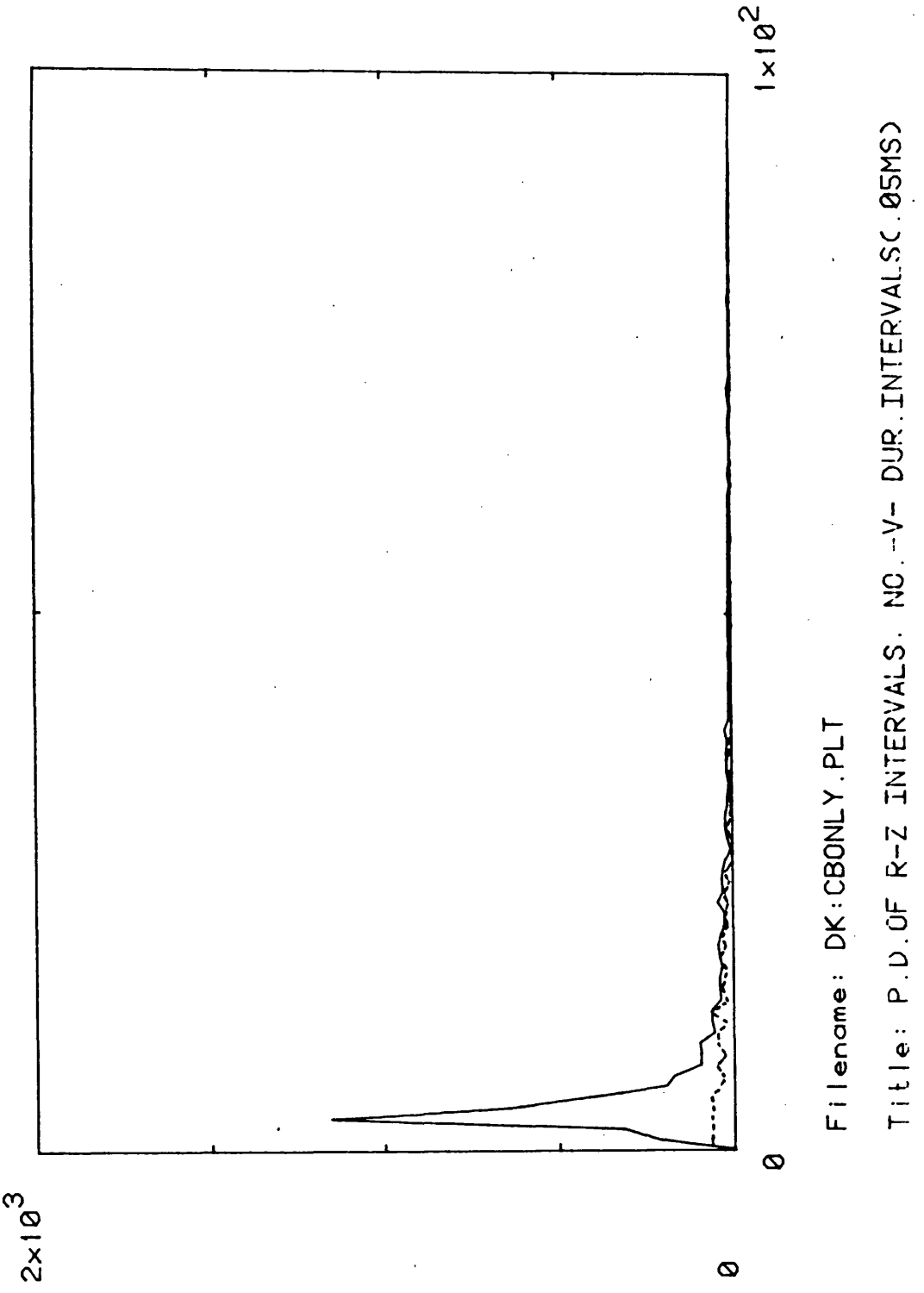


Figure 5.23

Comparison of probability distributions for speech file CBONLY.SPH

———— the original real-zero intervals

-------- symbols stored in the transmitter buffer

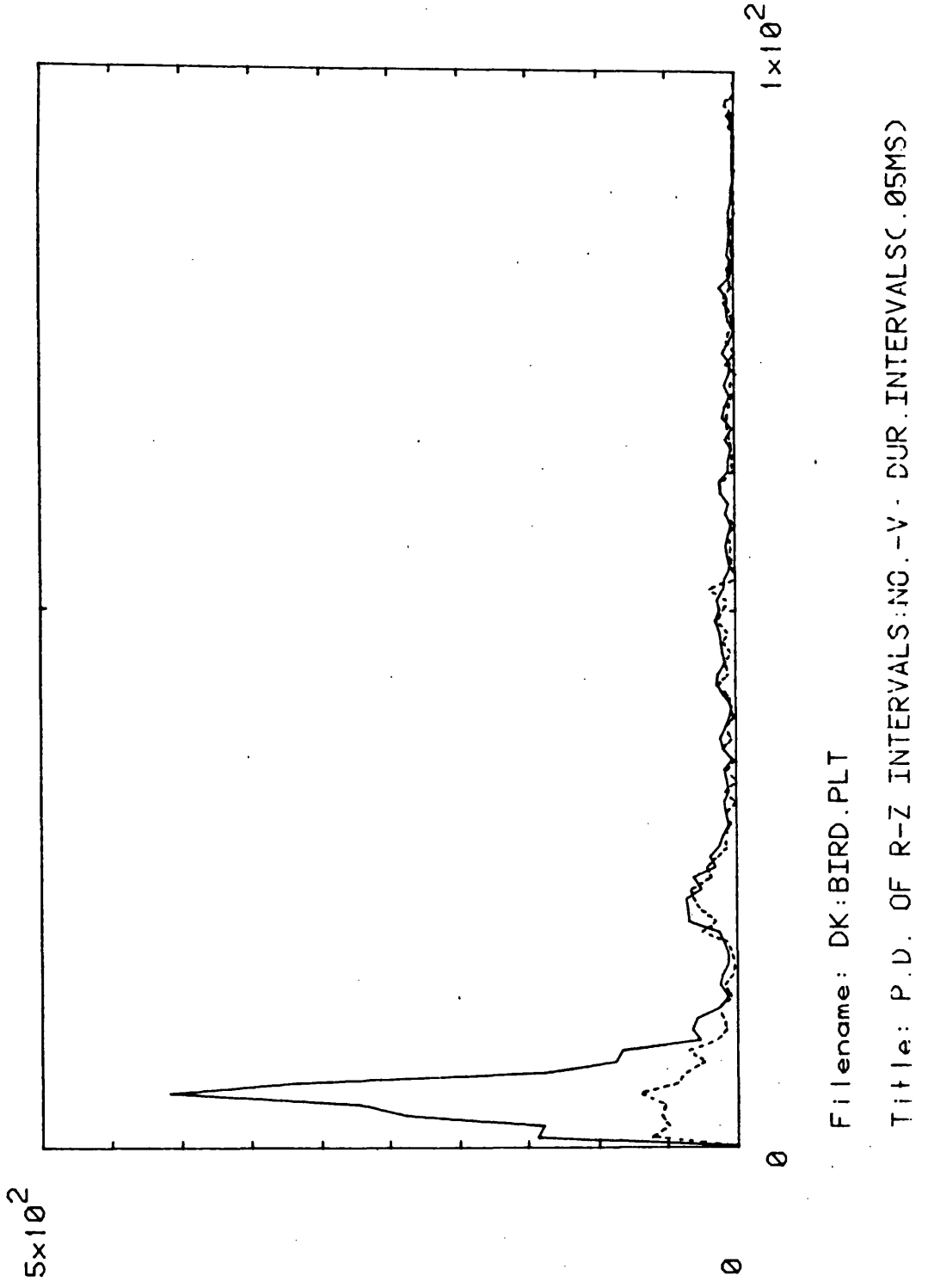


Figure 5.24

Comparison of probability distributions for speech file BIRD.SPH

——— the original real-zero intervals

------- symbols stored in the transmitter buffer

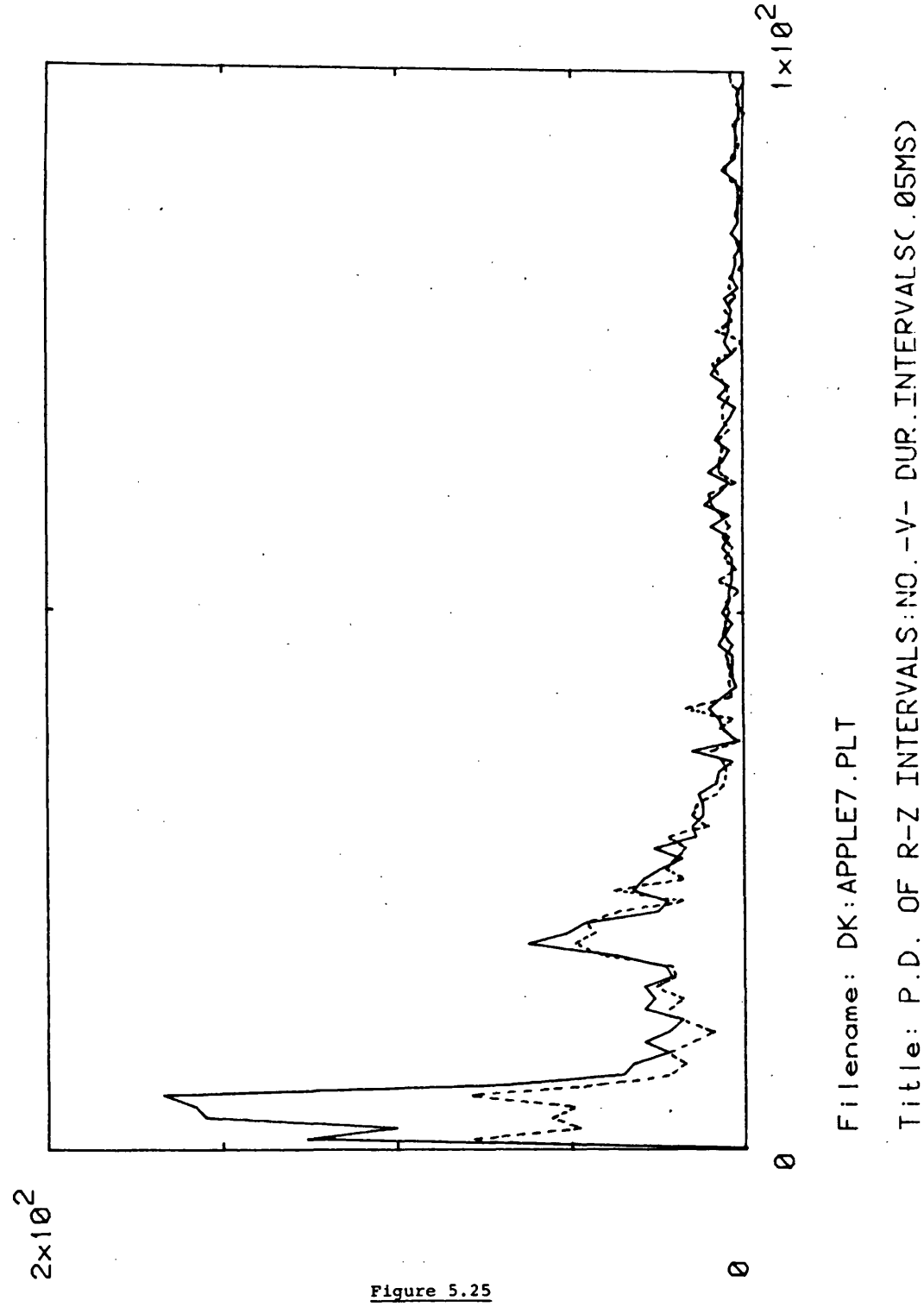


Figure 5.25

Comparison of probability distributions for speech file APPLE7.SPH

——— the original real-zero intervals

------- symbols stored in transmitter buffer

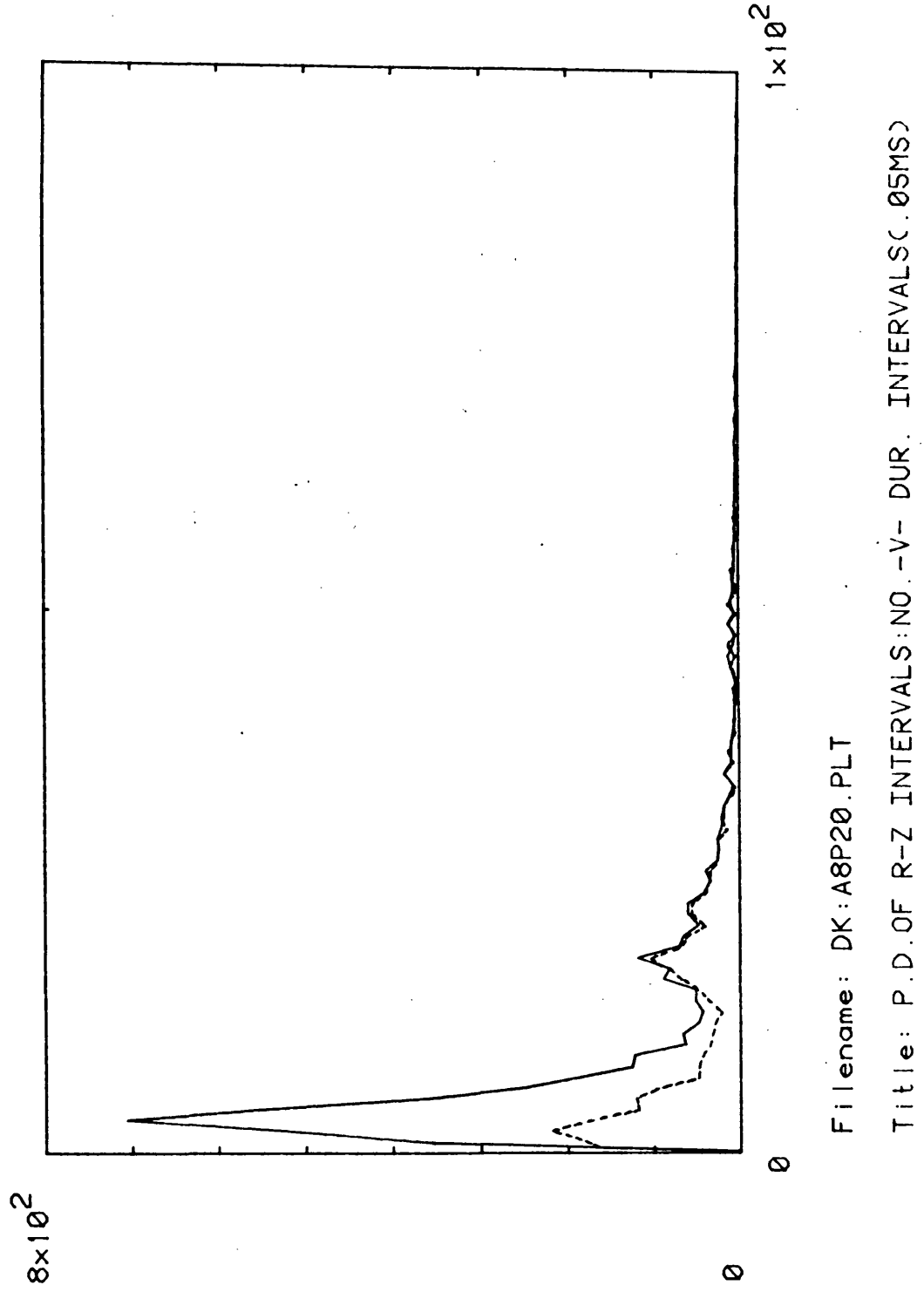


Figure 5.26 Comparison of probability distributions for a noisy speech file

Speech file A8P20 obtained by adding noise (SNR = 20dB) to speech file APPLE8.SPH.

——— The real-zero intervals of noisy file

------- Symbols stored in the transmitter buffer

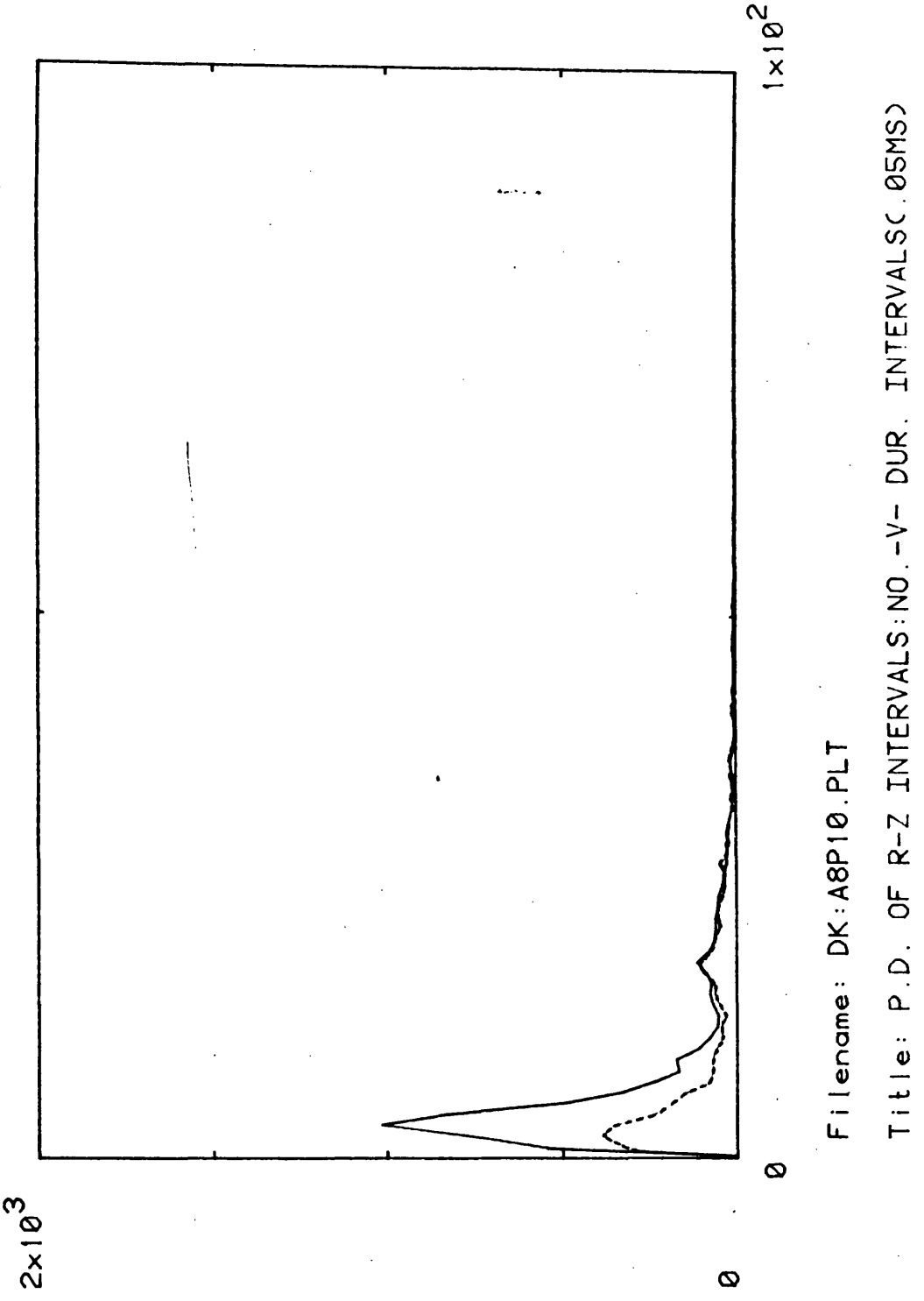


Figure 5.27 Comparison of probability distributions for a noisy speech file

Speech file A8P10 obtained by adding noise (SNR = 10dB) to speech file APPLE8.SPH.

———— The real-zero intervals of noisy file

-------- Symbols stored in the transmitter buffer

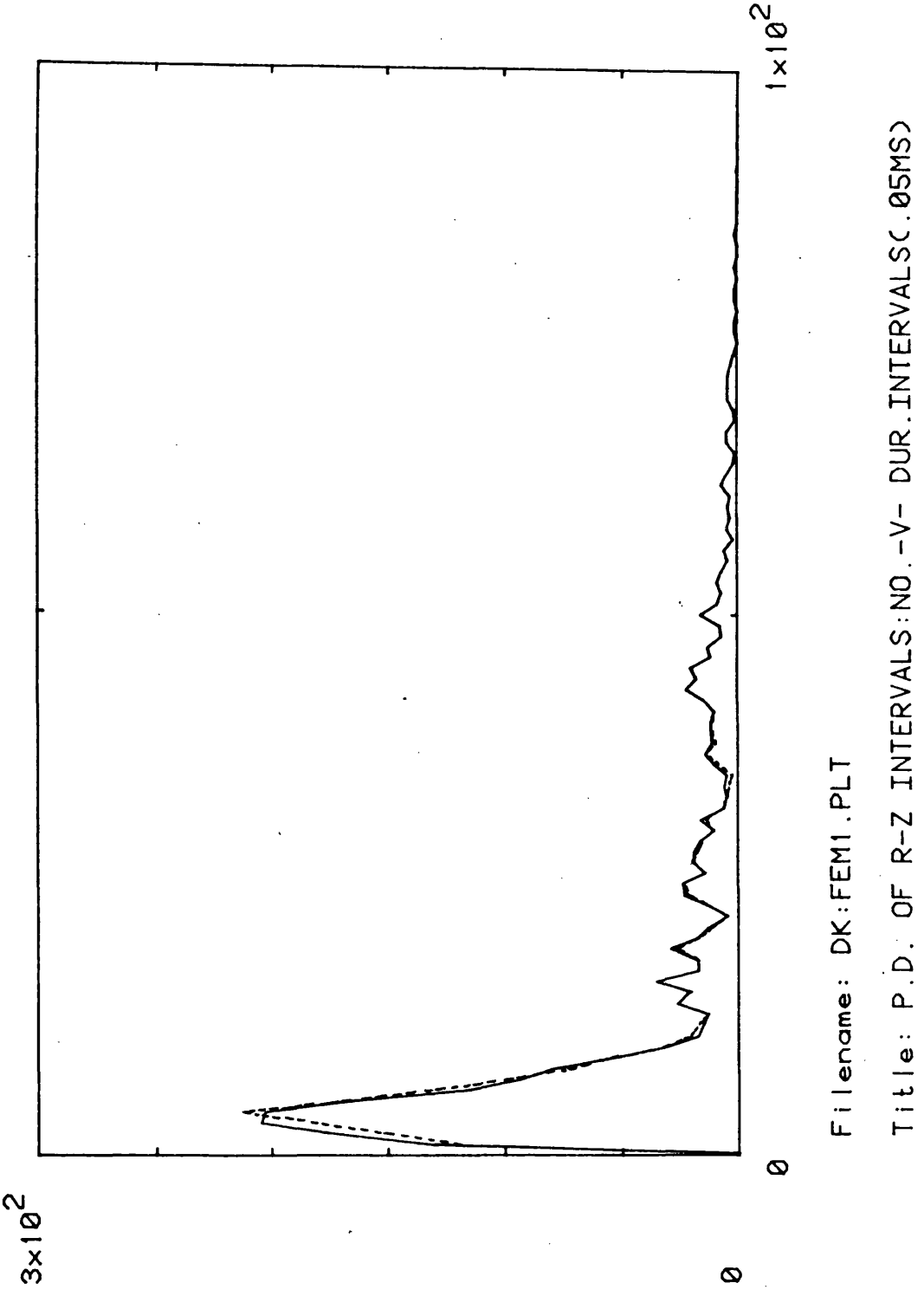


Figure 5.28

Comparison of probability distributions for speech file FEM1.SPH

——— the original real-zero intervals with silence replaced by zero amplitude samples

------- the reconstructed real-zero intervals

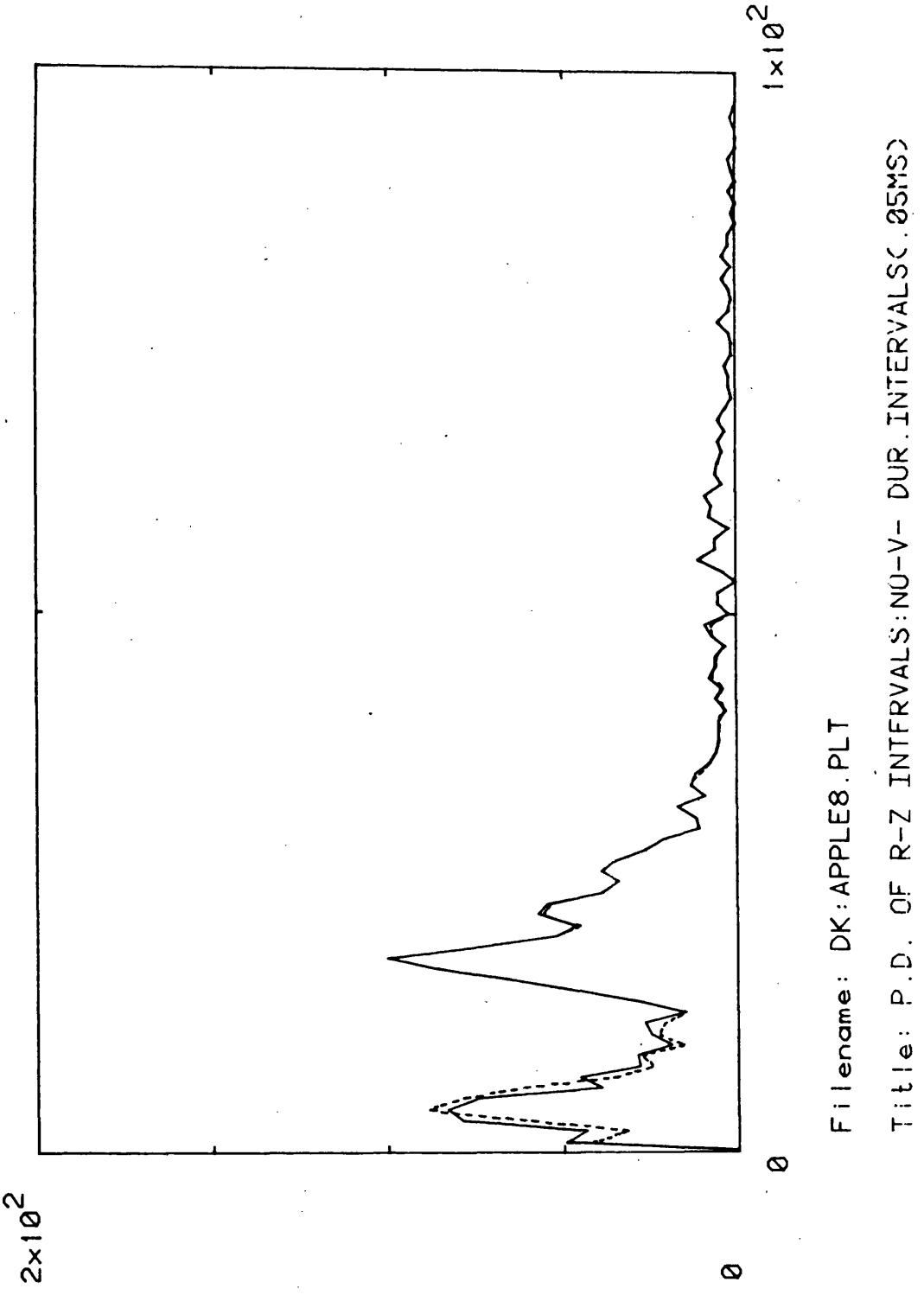


Figure 5.29

Comparison of probability distribution for speech file APPLE8.SPH

——— the original real-zero intervals with silence replaced by zero amplitude samples

------- the reconstructed real-zero intervals

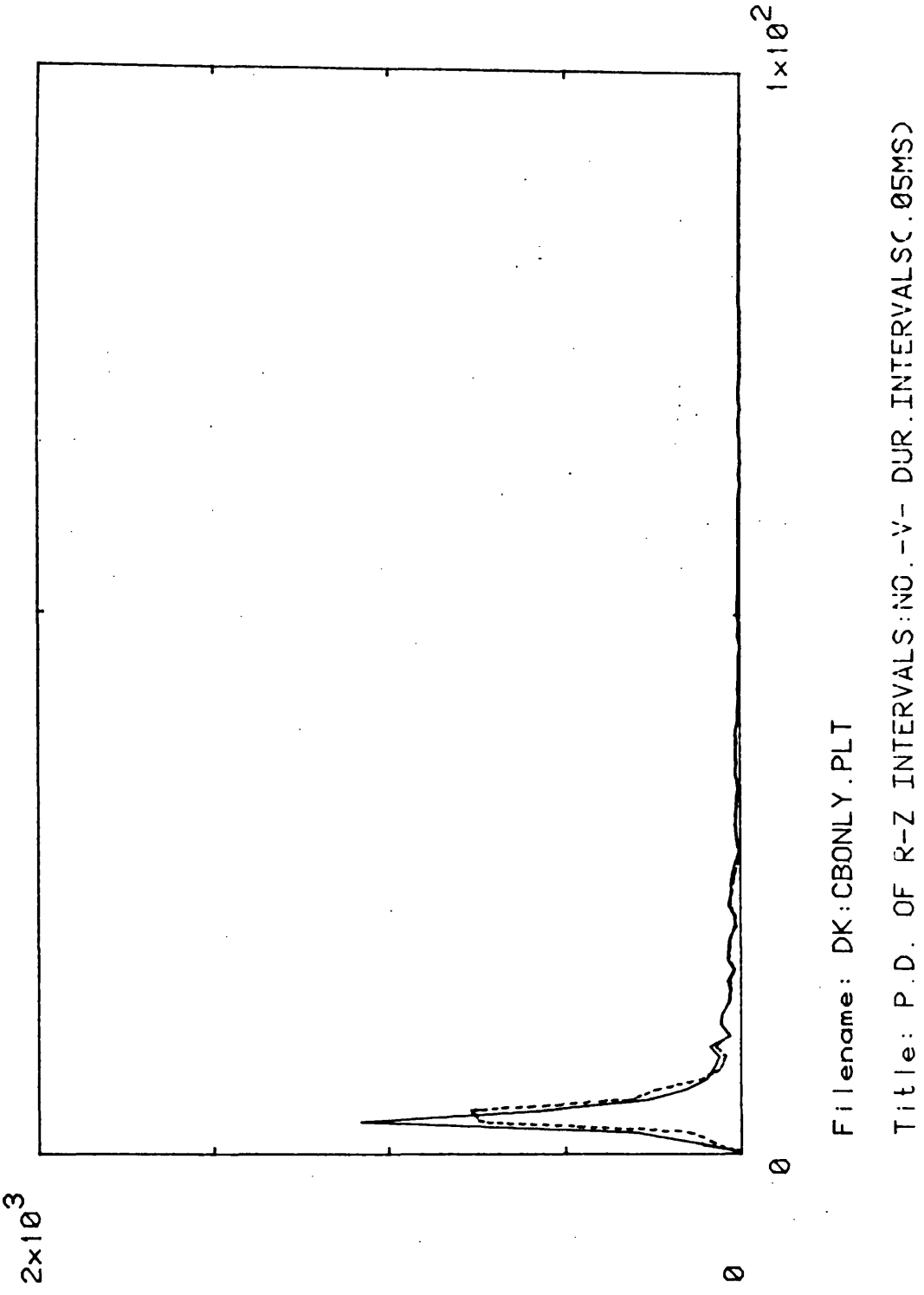


Figure 5.30

Comparison of probability distributions for speech file CBONLY.SPH

——— the original real-zero intervals with silence replaced by zero amplitude samples

------- the reconstructed real-zero intervals

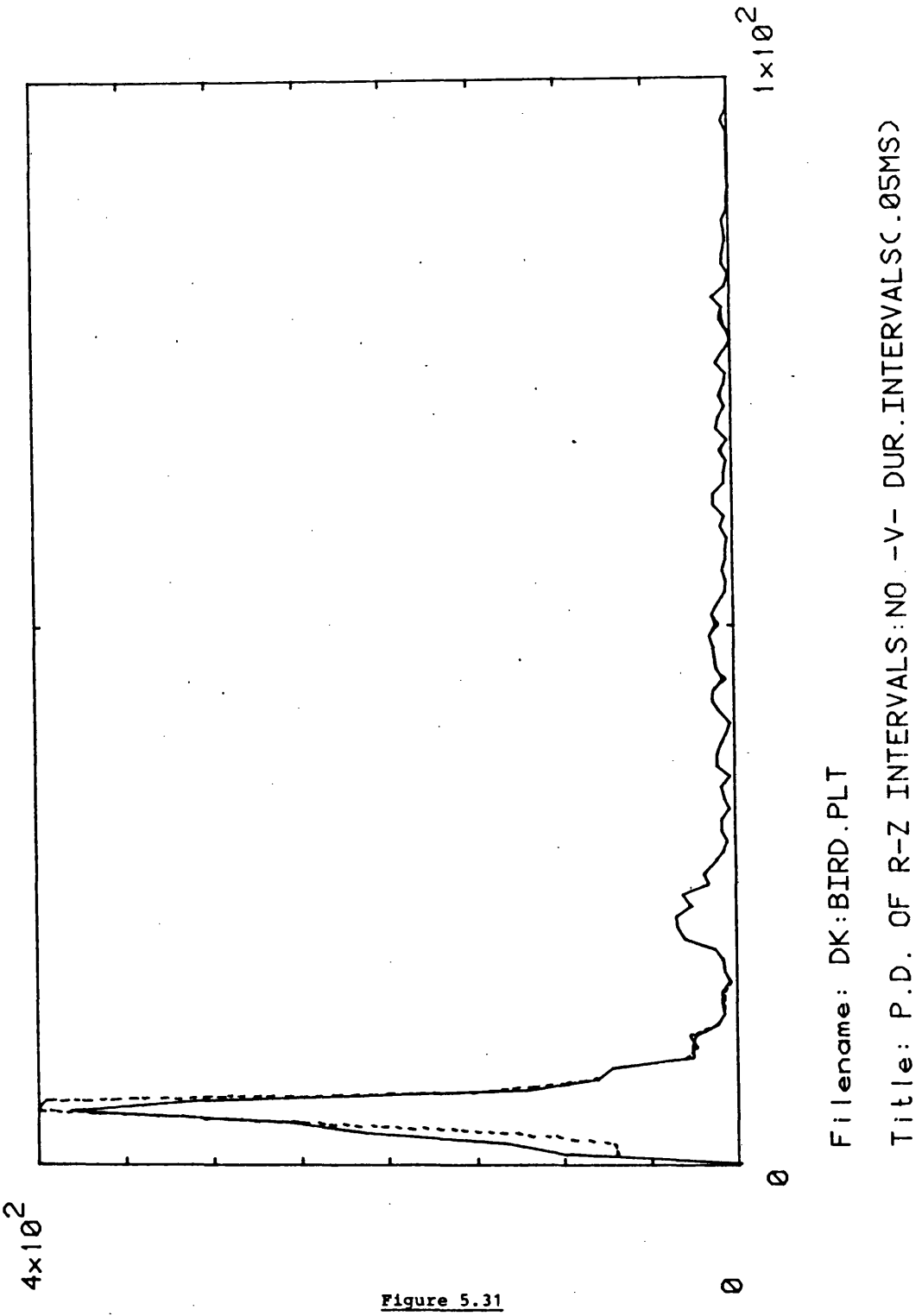


Figure 5.31

Comparison of probability distributions for speech file BIRD.SPH

——— the original real-zero intervals with silence replaced by zero amplitude samples

-------- the reconstructed real-zero intervals

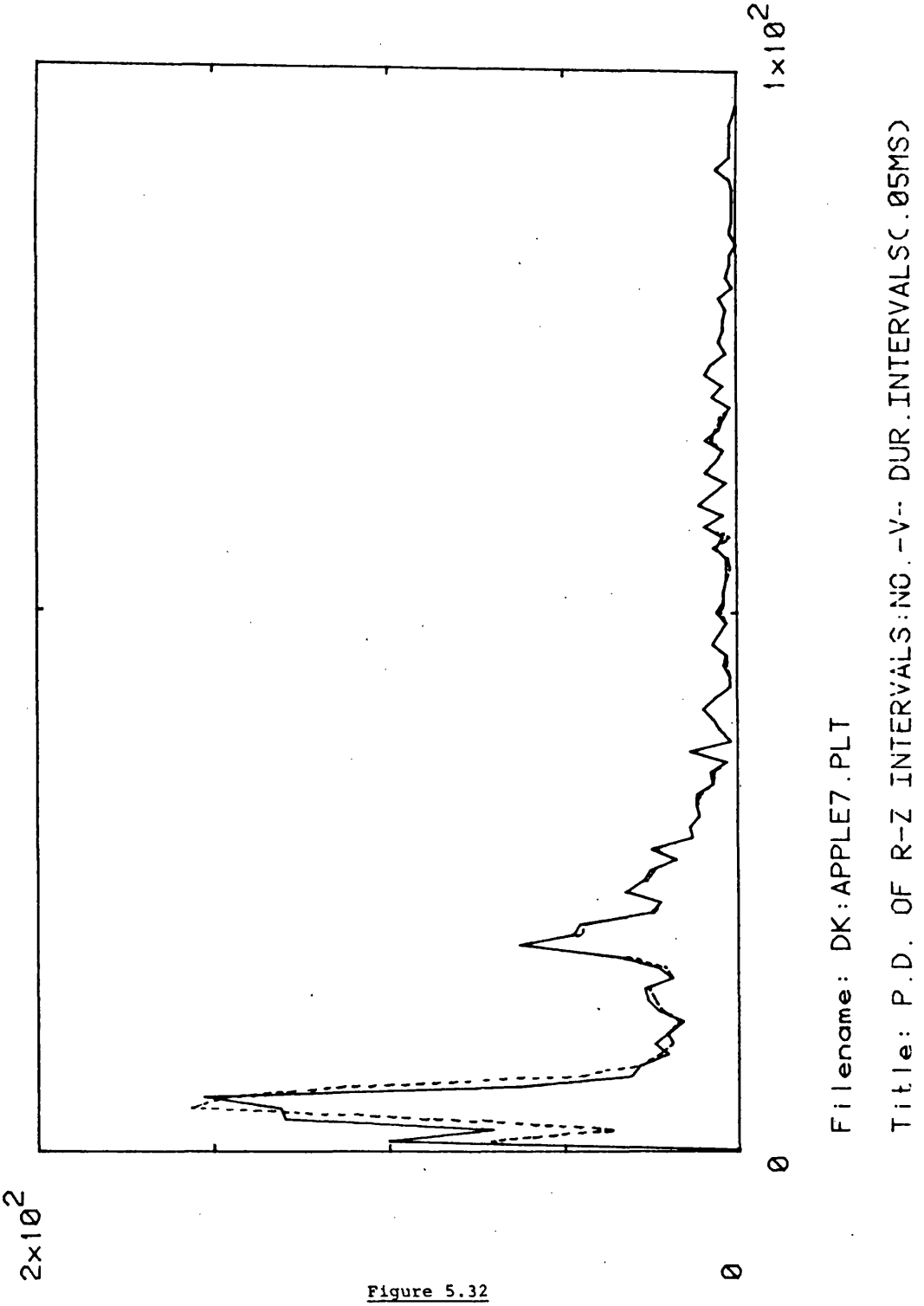


Figure 5.32

Comparison of probability distributions for speech file APPLE7.SPH

——— the original real-zero intervals with silence replaced by zero amplitude samples

------- the reconstructed real-zero intervals

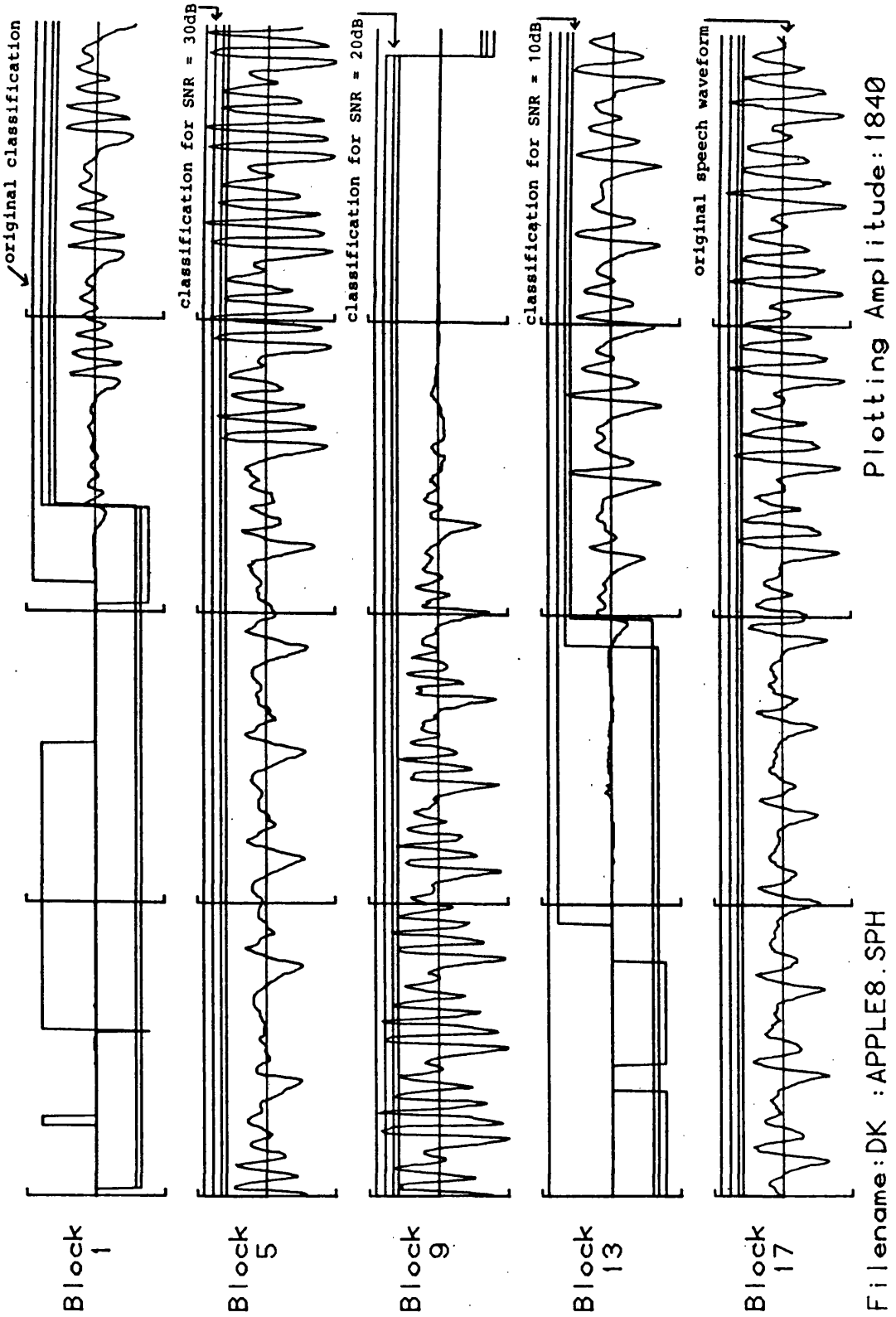


Figure 5.33

Illustrating the effect on classification of adding noise to the speech file APPLE8.SPH

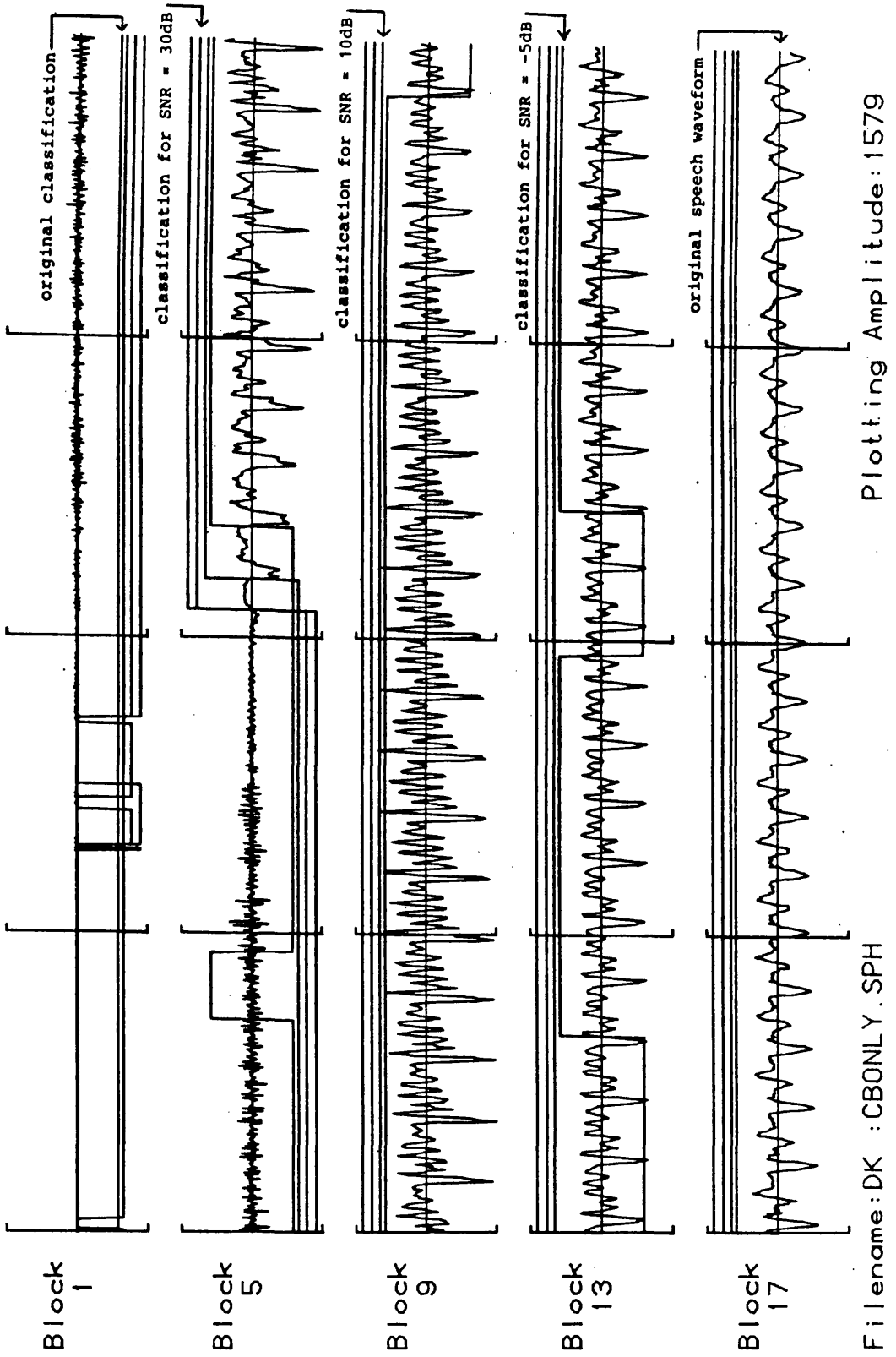


Figure 5.34

Illustrating the effect on classification of adding noise to the speech file CBONLY.SPH. Here, the classification algorithm is considered to fail at SNR = -5dB

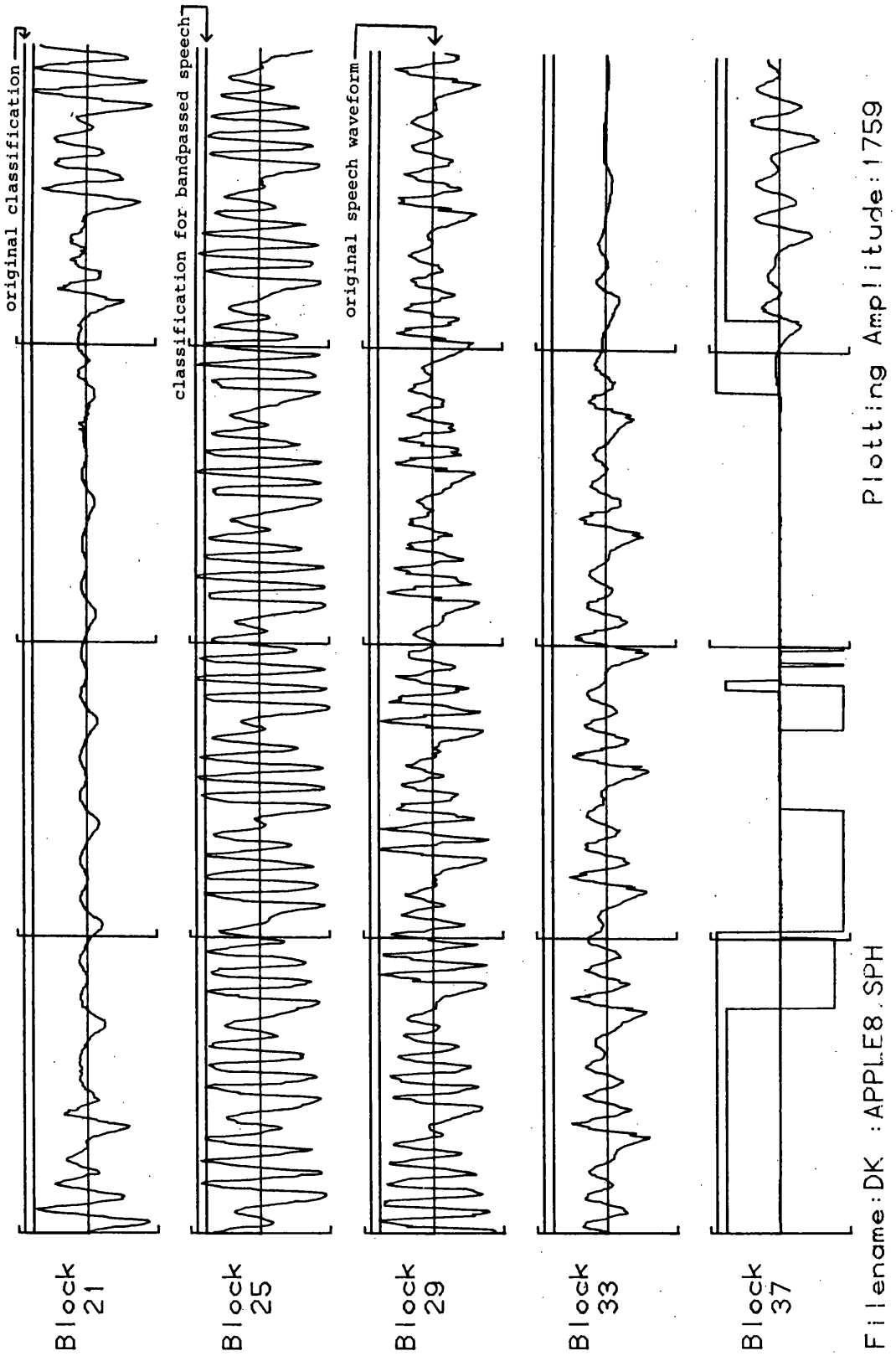


Figure 5.35

Illustrating the effect on classification of bandpass filtering the speech file APPLE8.SPH between 300Hz and 3400Hz

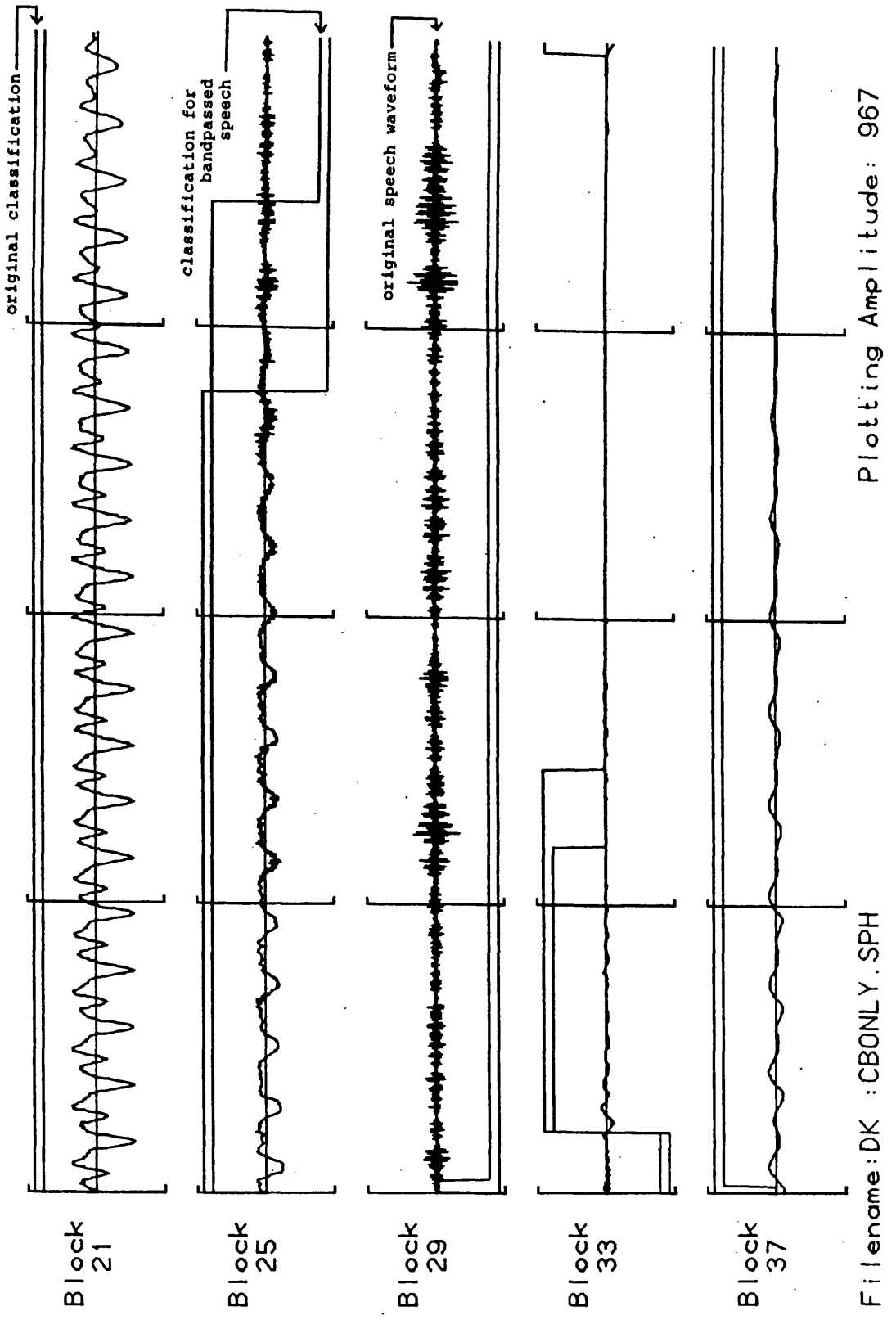


Figure 5.36

Illustrating the effect on classification of bandpass filtering the speech file CBONLY.SPH between 300Hz and 3400Hz

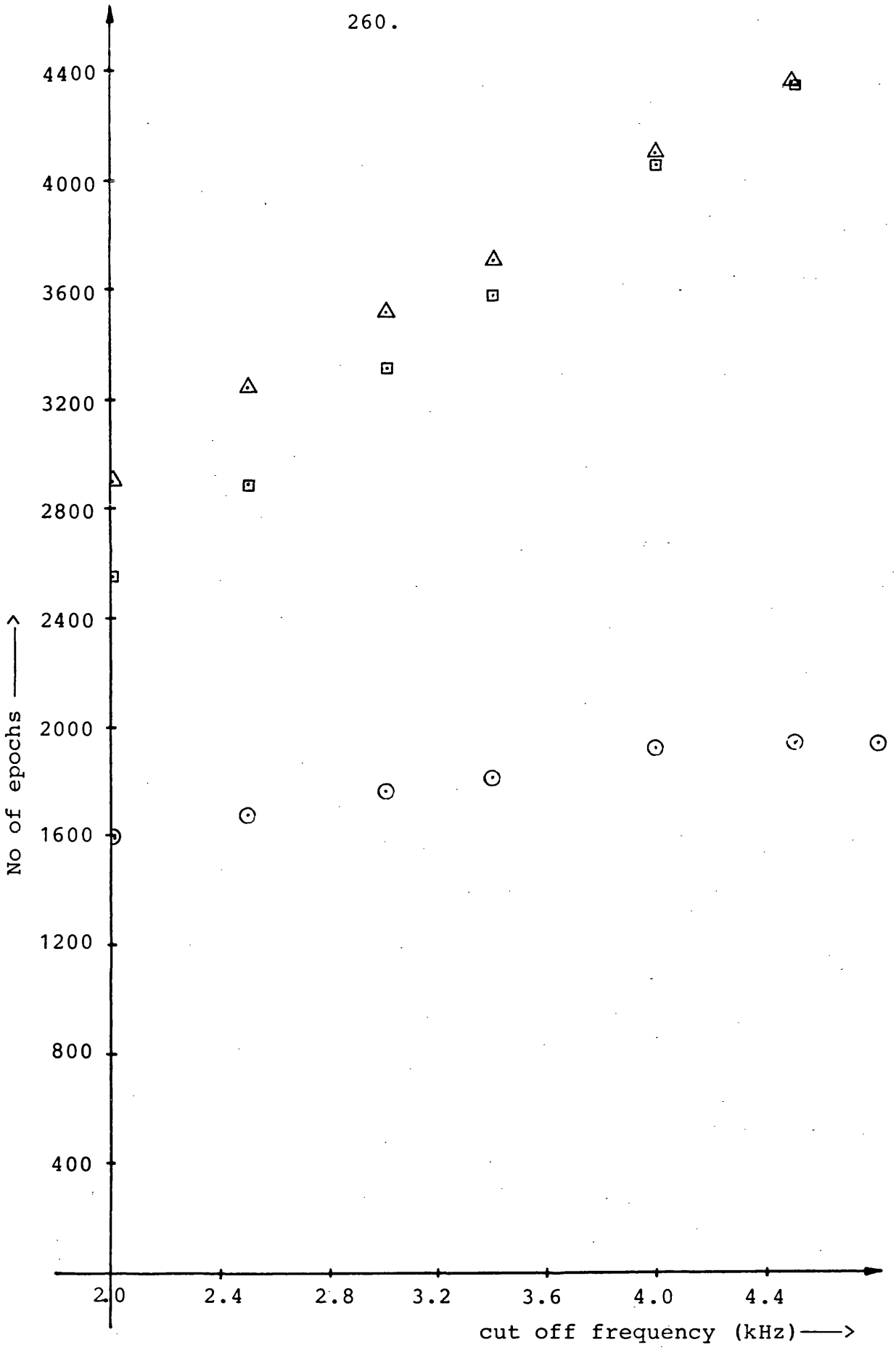


Figure 5.37 Effect on the epoch count of low pass filtering

△ 2nd order Butterworth, on CBONLY.SPH
⊡ 8th order Butterworth, on CBONLY.SPH
⊙ 2nd order Butterworth, on APPLE8.SPH

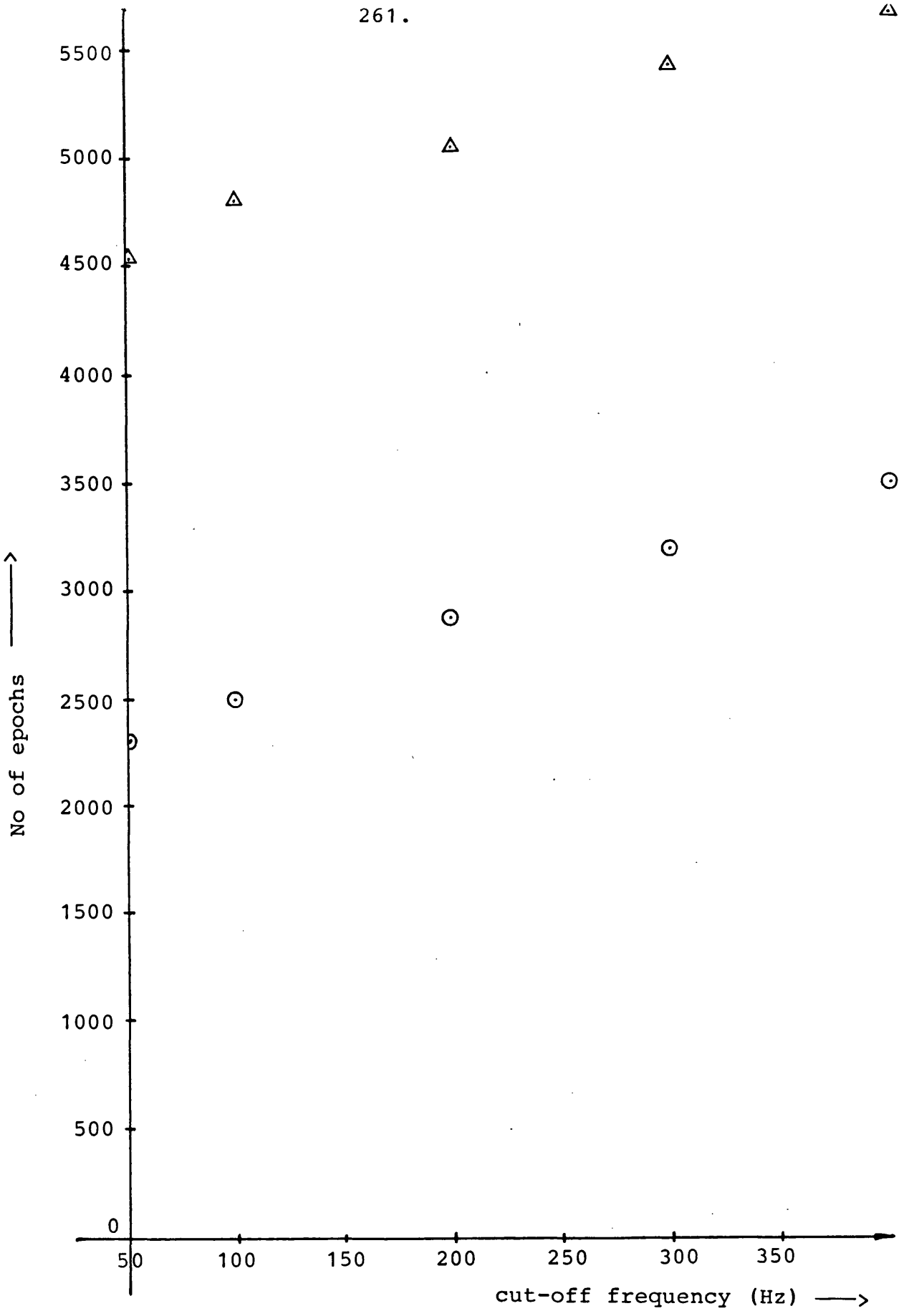


Figure 5.38 Effect on the epoch count of high pass filtering

△ 2nd order Butterworth, on CBONLY.SPH

⊙ 2nd order Butterworth, on APPLE8.SPH

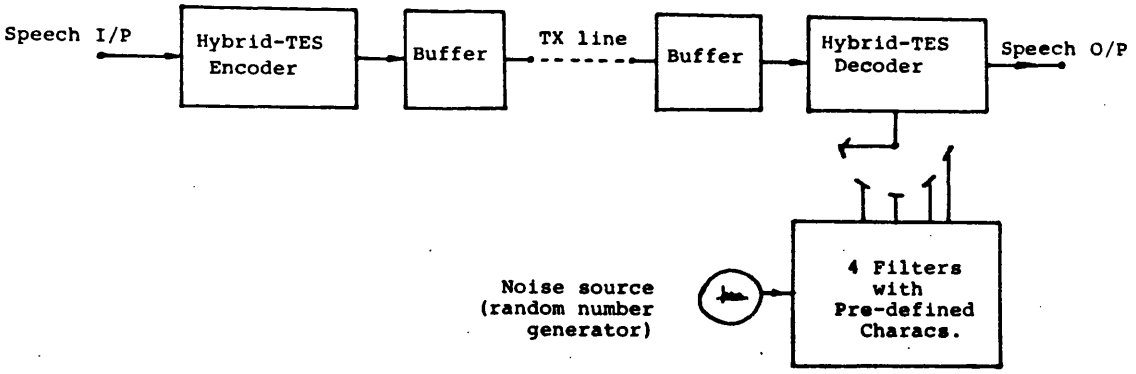


Figure 6.1 A Simplified Hybrid-TES Transmission System

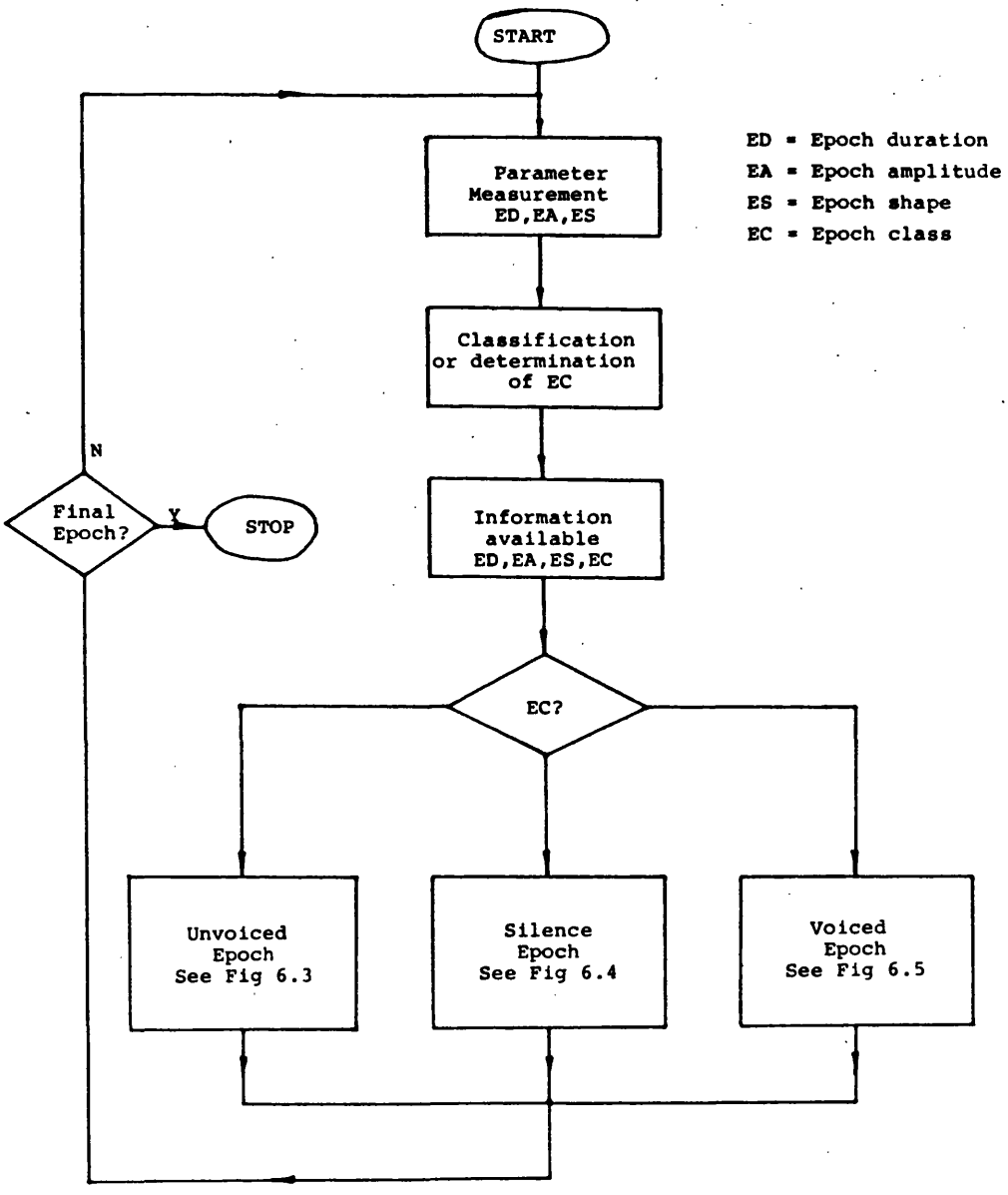


Figure 6.2 Flow diagram for Hybrid-TES encoding

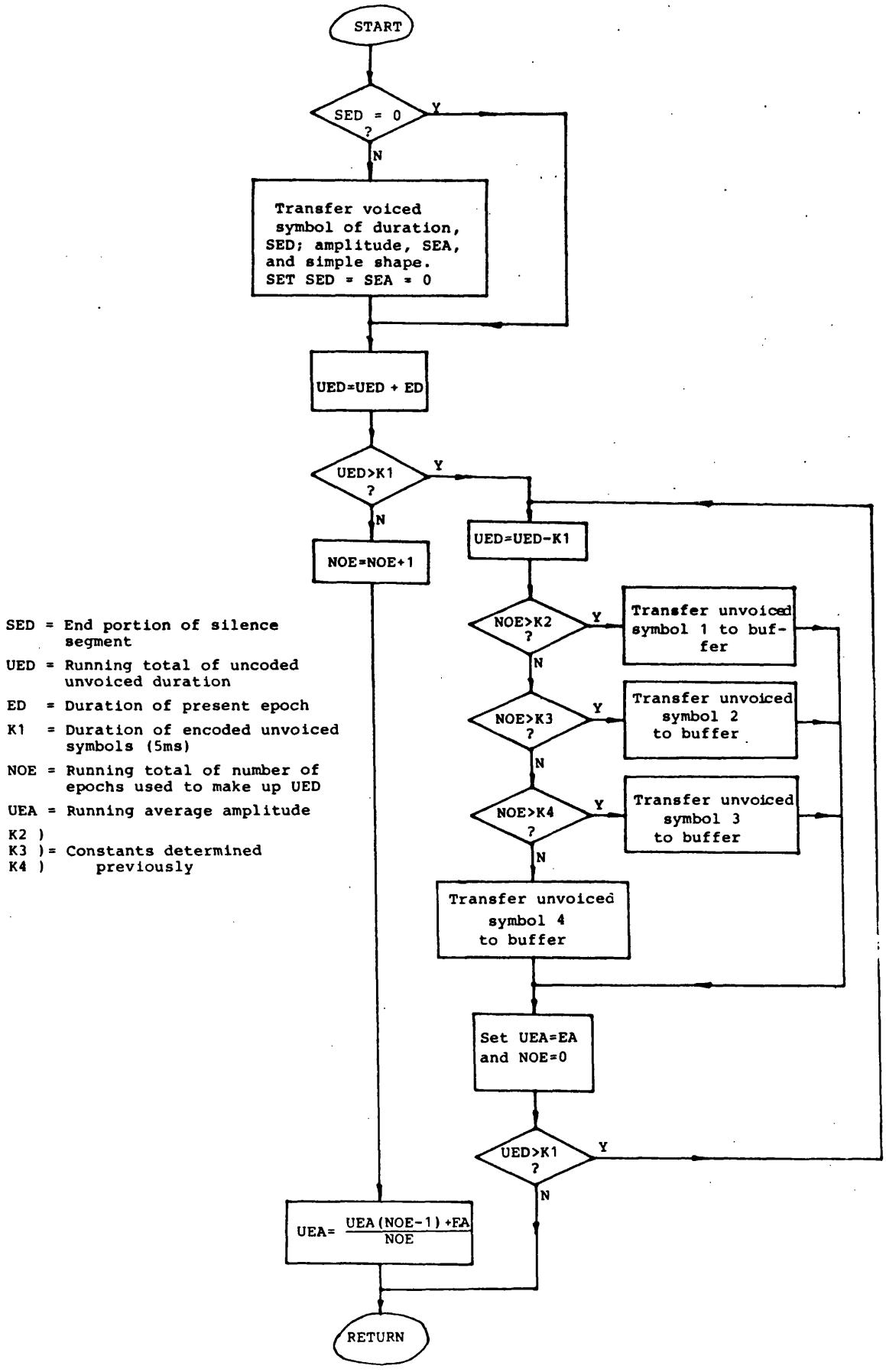


Figure 6.3 Flow diagram for processing unvoiced epochs

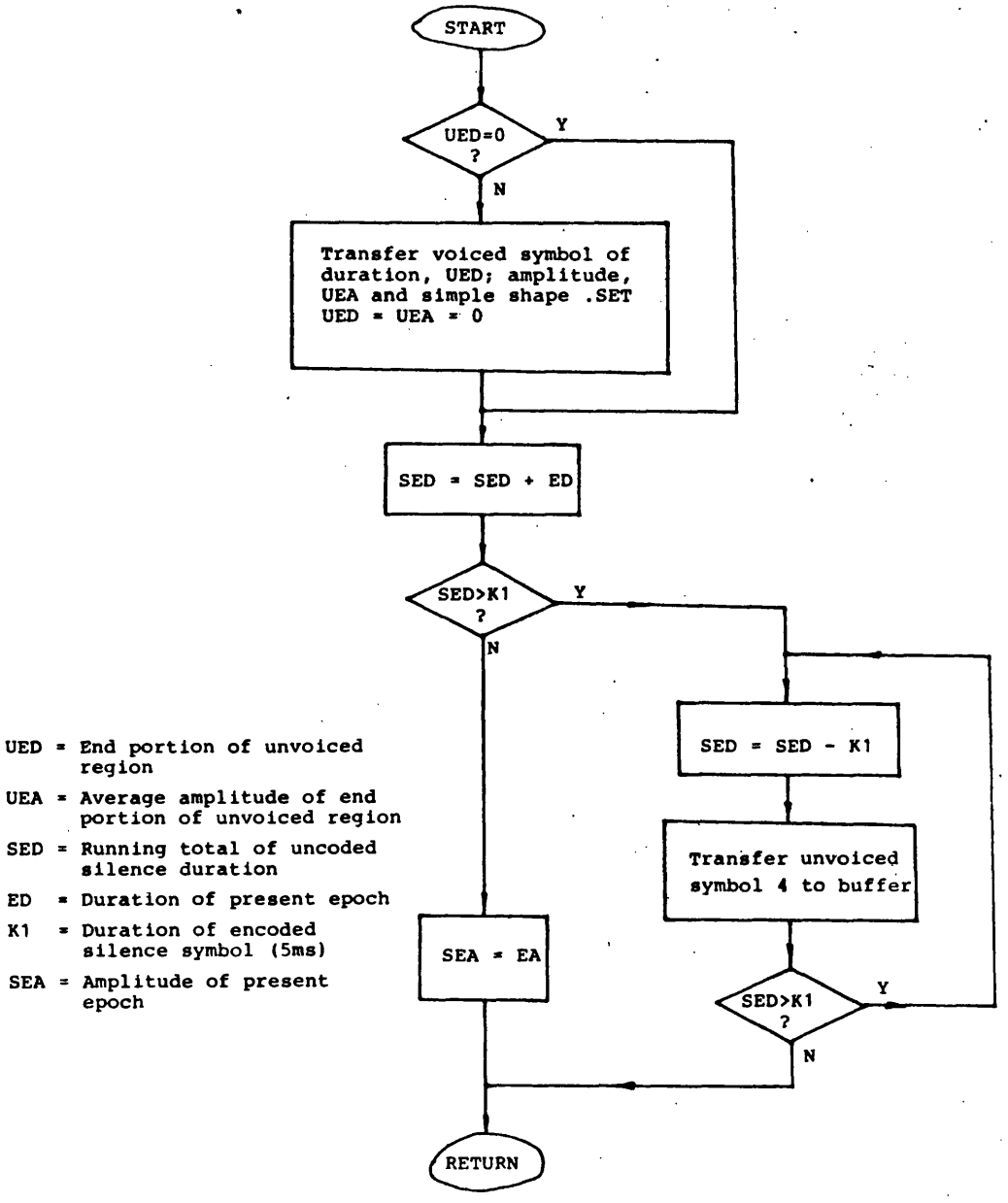


Figure 6.4 Flow diagram for processing silence epochs

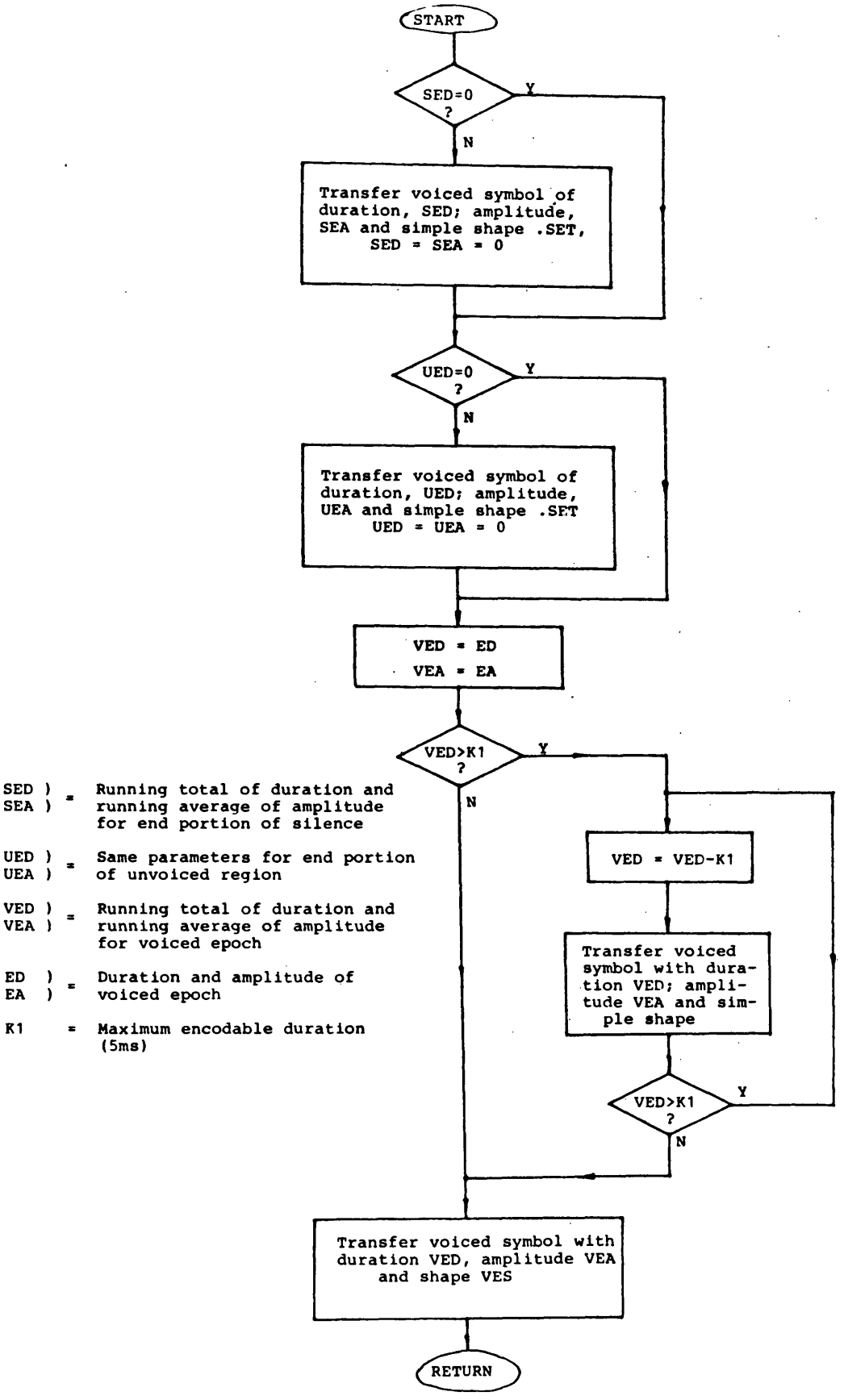


Figure 6.5 Flow diagram for processing voiced epochs

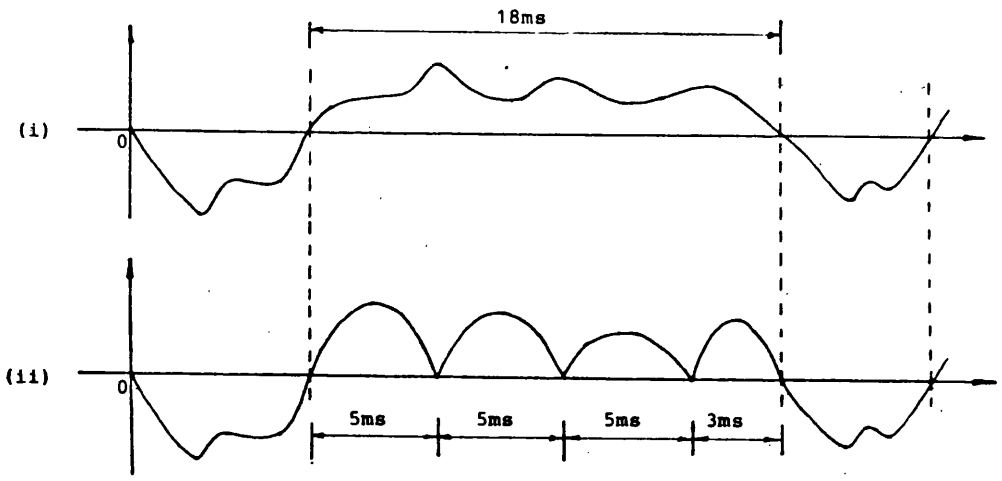


Figure 6.6 Voiced epochs in excess of 5ms (i) are encoded in segments of 5ms (ii)

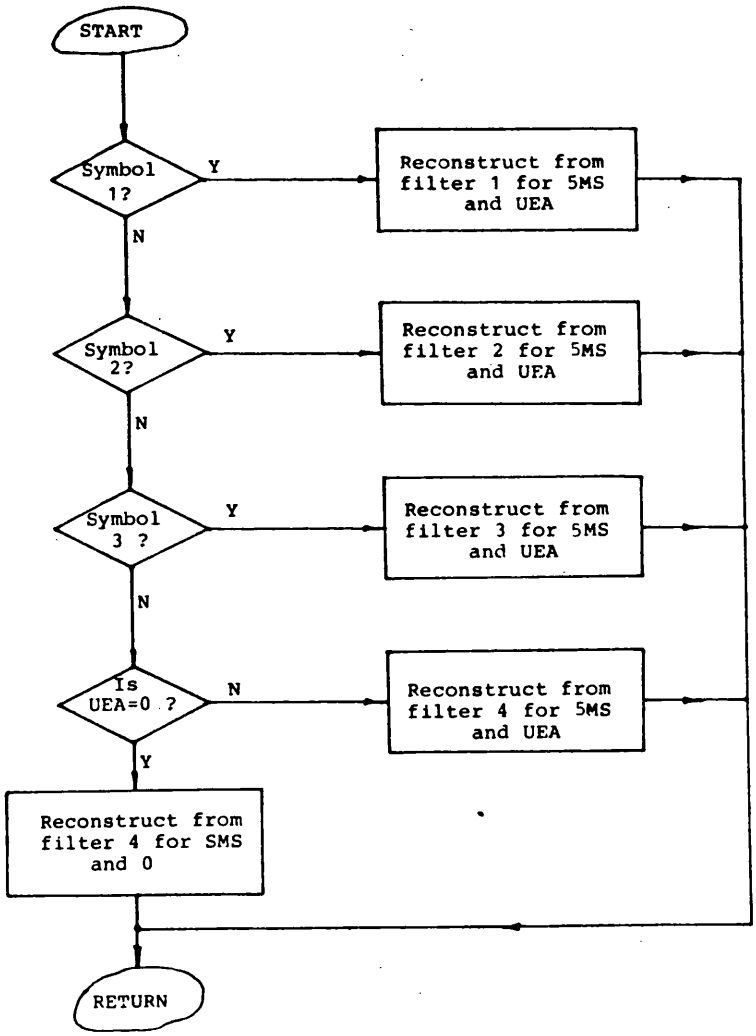


Figure 6.7 Reconstruction from unvoiced symbols

UEA = Unvoiced symbol amplitude

No of cycles	No of bytes	Label	Instruction		Comment
3	2		LDA A	EA	Load A with epoch amplitude
3	2		LDA B	ICLASS	Test previous class
3	2		CMP B	#0	
4	2		BLT	300	Branch to unvoiced state
4	2		BGT	200	Branch to voiced state
3	2	100	CMP A	T	
4	2		BLT	120	Indicates silence class
3	2		LDA B	KL	
3	2		CMP B	#1	
4	2		BLT	150	Indicates silence class
3	2		LDA B	ED	
3	2		CMP B	DUR	
4	2		BGT	130	Indicates voiced class
6	2	130	DEC	ICLASS	Change ICLASS to unvoiced
4	2		STA A	PAV	Set PAV = EA
3	2		LDA B	#0	
4	2		STA B	KL	Set KL = 0
4	2		BRA	400	Indicates unvoiced class
6	2		INC	ICLASS	Change ICLASS to voiced
4	2		STA A	PAV	Set PAV = EA
3	2		LDA B	#0	
4	2		STA B	KL	Set KL = 0
4	2		BRA	600	Indicates voiced class
		120	LDA B	#0	
			STA B	KL	Set KL = 0
		150	BRA	500	Indicates silence class

Time to detect: silence class = 35 or 38 machine cycles
unvoiced class = 65 machine cycles
voiced class = 65 machine cycles

No of cycles	No of bytes	Label	Instruction		Comment
4	2	200	SUB A	PAV	Determine approximate PAV
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
3	2		ADD A	PAV	
4	2		STA A	PAV	
3	2		CMP A	T	
4	2		BLT	210	Indicates silence class
3	2		LDA B	ED	
3	2		CMP B	DUR	
4	2		BGT	220	Indicates voiced class
3	2		LDA B	NB	
3	2		CMP B	LIM	
4	2		BLT	230	Indicates voiced class
3	2		LDA B	#-1	
4	2		STA B	ICLASS	Change ICLASS to unvoiced
4	2		BRA	400	Indicates unvoiced class
6	2	210	DEC	ICLASS	Change ICLASS to silence
4	2		BRA	500	Indicates silence class
3	2	220	LDA B	#0	
4	2		STA B	NB	Set NB = 0
4	2		BRA	600	Indicates voiced class
6	2	230	INC	NB	
			BRA	600	Indicates voiced class

Time to detect: Voiced class = 66 or 75 machine cycles
unvoiced class = 76 machine cycles
silence class = 45 machine cycles

No of cycles	No of bytes	Label	Instruction		Comment
4	2	300	SUB A	PAV	Determine approximate PAV
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
3	2		ADD A	PAV	
4	2		STA A	PAV	
3	2		CMP A	T	
4	2		BLT	310	Indicates silence class
3	2		LDA B	ED	
3	2		CMP B	DUR	
4	2		BGT	320	Indicates possibility of unvoiced class
4	2		BRA	400	Indicates unvoiced class
6	2	310	INC	ICLASS	Change ICLASS to silence
4	2		BRA	500	
6	2	320	INC	KL	
3	2		LDA B	KL	
3	2		CMP B	#1	
4	2		BLE	400	Indicates unvoiced class
2	1		DEC B		
4	2		STA B	ICLASS	Change ICLASS to voiced
2	1		DEC B		
4	2		STA B	NB	Set NB = 0
4	2		BRA	600	Indicates voiced class

Time to detect: unvoiced class = 55 or 67 machine cycles
silence class = 51 machine cycles
voiced class = 83 machine cycles

Figure 6.8 Simple source program for the classification algorithm

Total number of bytes = 142

UNVOICED CLASS

No of cycles	No of bytes	Label	Instruction		Comment
3	2	400	LDA B	SED	Load B with silence duration
3	2		CMP B	#0	Indicates zero silence
4	2		BEQ	410	
3	2		LDA B	SYMBOL	Store voiced symbol in buffer
4	2		STA B	LINE	
3	2		LDA A	#0	Reset SED and SEA
4	2		STA A	SED	
4	2		STA A	SEA	
3	2	410	LDA B	UED	
3	2		ADD B	ED	
3	2		CMP B	#K1	Indicates unvoiced duration less than segment K1
4	2		BLT	420	
4	2	460	SUB B	#K1	Decision as to which unvoiced symbol to store in transmit buffer
3	2		LDA A	NOE	
3	2		CMP A	#K2	
4	2		BGT	430	
3	2		CMP A	#K3	
4	2		BGT	440	
3	2		CMP A	#K4	
4	2		BGT	450	
3	2		LDA A	SYMBOL4	
4	2	480	STA A	LINE	Store relevant symbol in buffer
3	2		LDA A	EA	Set UEA = EA
3	2		STA A	UEA	
3	2		LDA A	#0	Set NOE = 0
4	2		STA A	NOE	
3	2		CMP B	K1	Indicates remaining duration less than segment
4	2		BGT	460	
4	2	470	STA B	UED	Encoding completed
4	2		BRA	800	
3	2	630	LDA A	SYMBOL1	
4	2		BRA	480	
3	2	640	LDA A	SYMBOL2	
4	2		BRA	480	
3	2	450	LDA A	SYMBOL 3	
4	2		BRA	480	
3	2	420	LDA A	NOE	Increment NOE
2	1		INC A		
4	2		STA A	NOE	
2	1		ASR A		Determination of approximate average unvoiced amplitude
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
2	1		ASR A		
3	2		ADD A	UEA	
4	2		STA A	UEA	
4	2		BRA	470	

Maximum execution time : 105 machine cycles

SILENCE CLASS

No of cycles	No of bytes	Label	Instruction		Comment
3	2	500	LDA B	UED	Load B with unvoiced duration
3	2		CMP B	#0	Indicates no unvoiced duration
4	2		BEQ	510	
3	2		LDA B	SYMBOL	Store voiced symbol in buffer
4	2		STA B	LINE	
3	2		LDA A	#0	Reset UED and UEA
4	2		STA A	UED	
4	2		STA A	UEA	
3	2	510	LDA B	SED	Indicates silence duration less than segment K1
3	2		ADD B	ED	
3	2		CMP B	#K1	
4	2		BLT	520	
4	2	530	SUB B	#K1	Store symbol 4 in buffer
3	2		LDA A	SYMBOL4	
4	2		STA A	LINE	
3	2		CMP B	#K1	Indicates remaining silence duration greater than segment K1
4	2		BGT	530	
4	2	540	STA B	SED	Encoding completed
4	2		BRA	800	
3	2	520	LDA A	EA	Set SEA = EA
4	2		STA A	SEA	
4	2		BRA	540	

Maximum execution time : 67 machine cycles

VOICED CLASS

No of cycles	No of bytes	Label	Instruction		Comment
3	2	600	LDA B	SED	Load B with silence duration
3	2		CMP B	#0	Indicates no silence duration
4	2		BEQ	610	
3	2		LDA B	SYMBOL	Store voiced symbol in buffer
4	2		STA B	LINE	
3	2		LDA A	#0	Reset SED and SEA
4	2		STA A	SED	
4	2		STA A	SEA	
3	2	610	LDA B	UED	Load B with unvoiced duration
3	2		CMP B	#0	Indicates no unvoiced duration
4	2		BEQ	620	
3	2		LDA B	SYMBOL	Store voiced symbol in buffer
3	2		STA B	LINE	
4	2		STA A	UED	Reset UED and UEA
4	2		STA A	UEA	
3	2	620	LDA A	ED	Indicates voiced duration less than segment K1
3	2		CMP A	#K1	
4	2		BLT	630	
4	2	640	SUB A	#K1	Store voiced symbol in buffer
3	2		LDA B	SYMBOL	
4	2		STA B	LINE	
3	2		CMP A	#K1	Indicates remaining ED greater than segment K1
4	2		BGT	640	
3	2	630	LDA B	SYMBOL	Store voiced symbol in buffer
4	2		STA B	LINE	
4	2		BRA	800	Encoding completed

Maximum execution time : 91 machine cycles

Figure 6.9 Simple source program for the encoding algorithm

Total number of bytes = 184

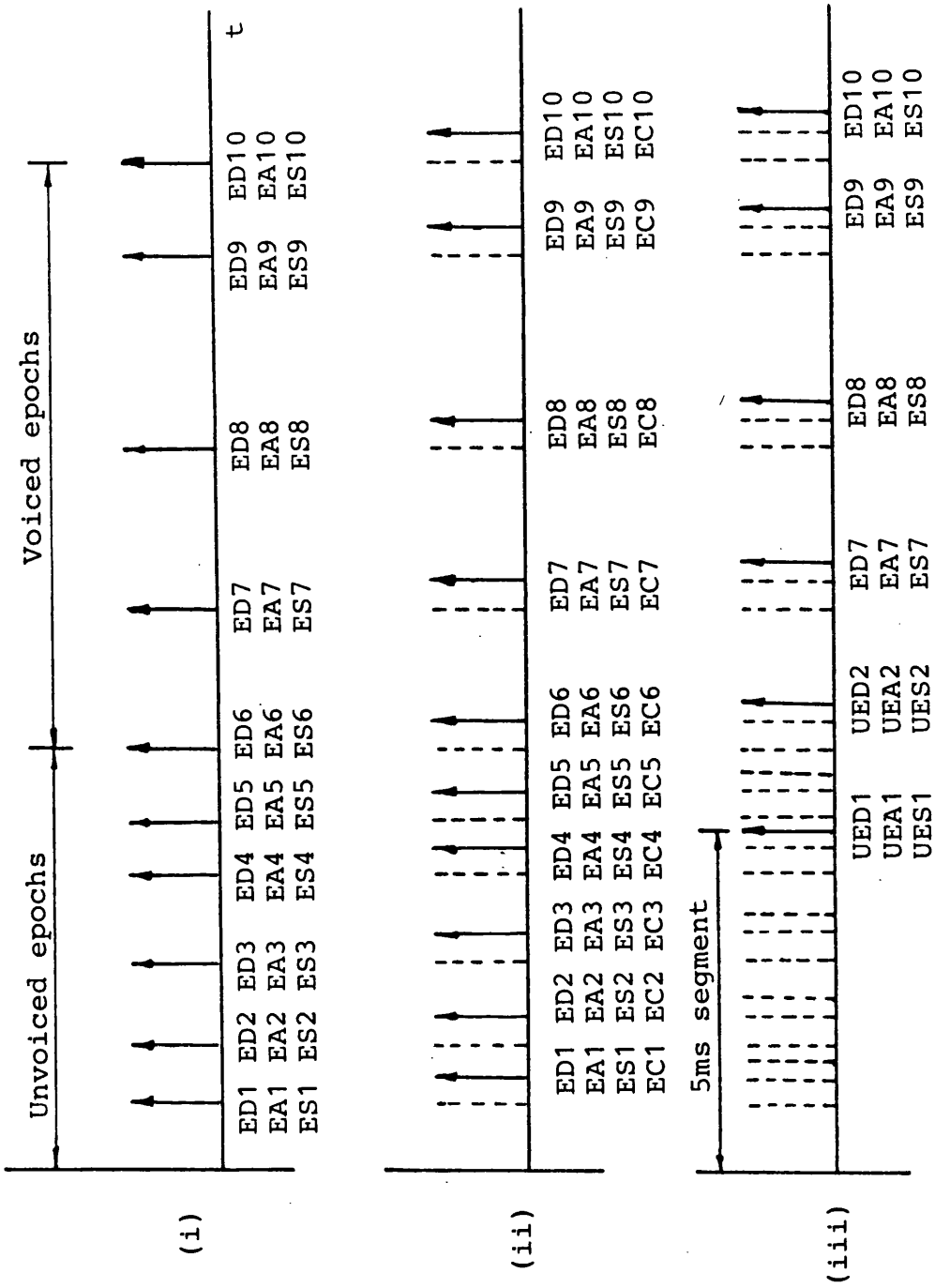


Figure 6.10 A typical timing chart for hybrid-TES

(i) Information available after epoch parameter measurement

(ii) Information available after classification

(iii) Information available after encoding

EA = Epoch amplitude; ED = epoch duration;
ES = epoch shape; EC = epoch class;
UED = unvoiced symbol duration;
UEA = unvoiced symbol amplitude; and
UES = unvoiced symbol shape

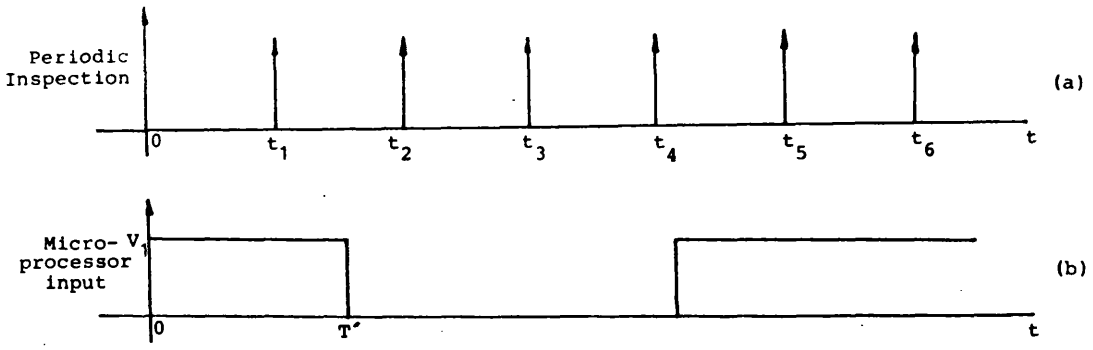


Figure A.1 The Flag Inspection Procedure

Change in microprocessor input will not be detected until the following inspection (eg: t_2 and t_5).

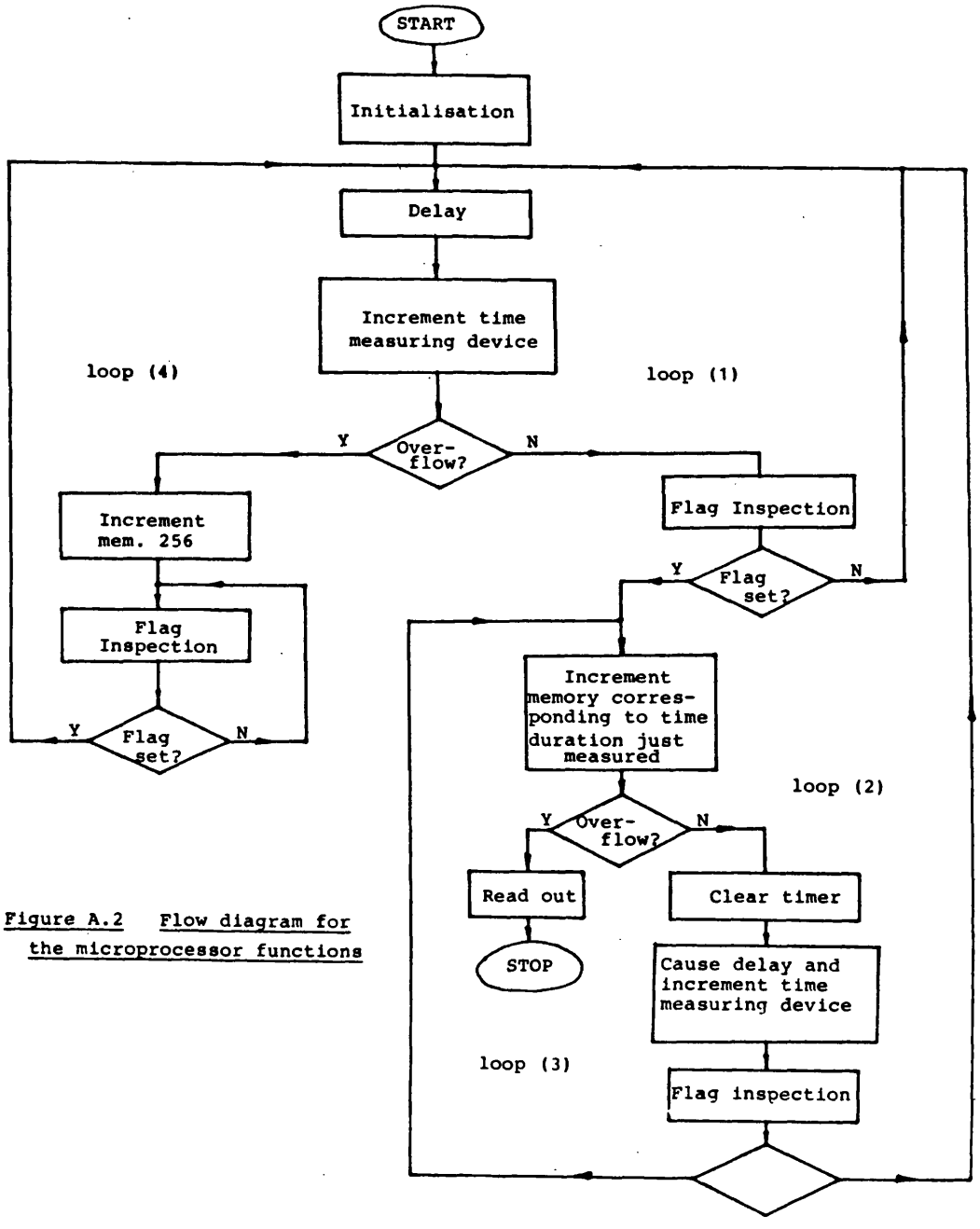


Figure A.2 Flow diagram for the microprocessor functions

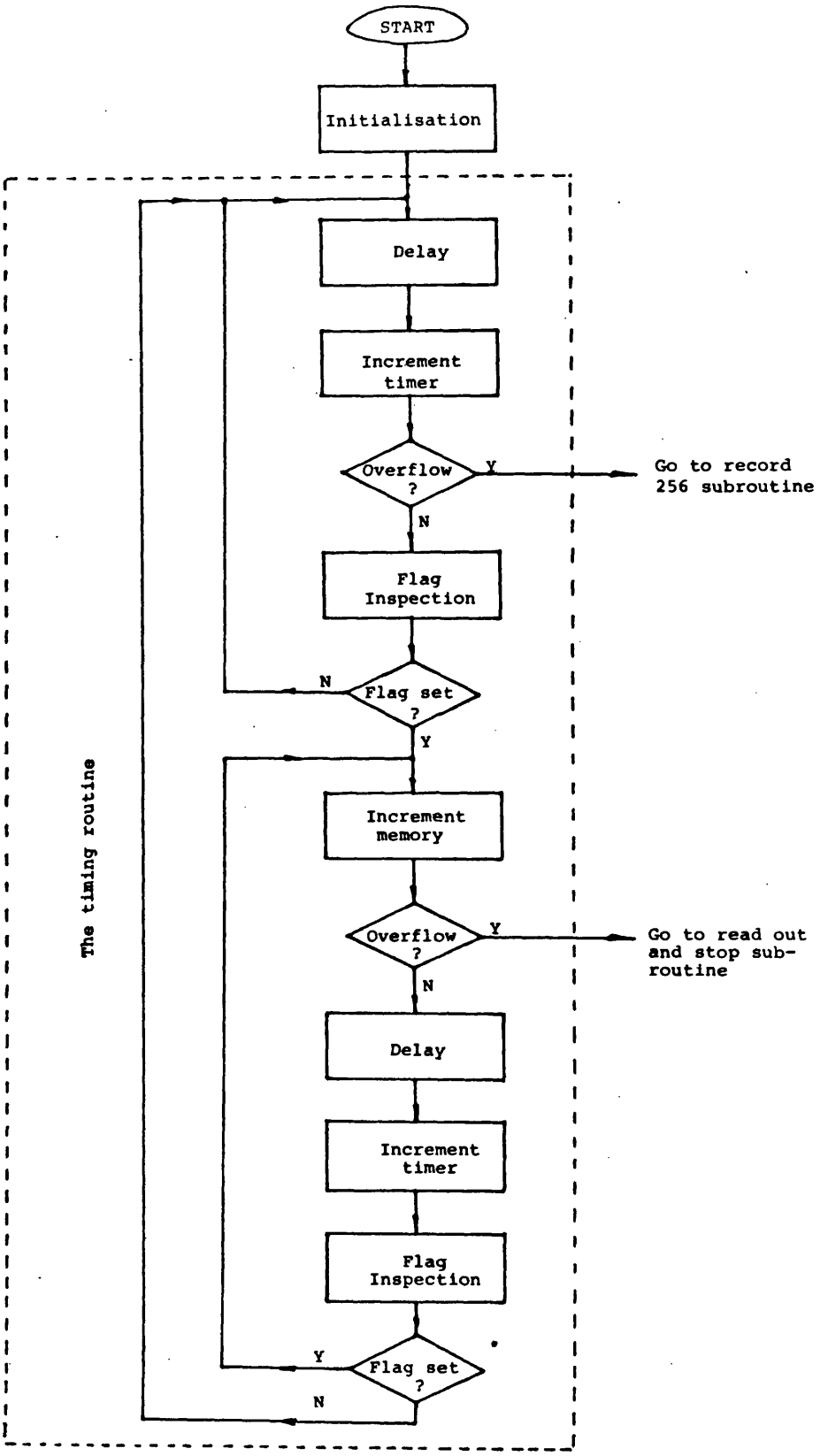


Figure A.3 The microprocessor functions subdivided

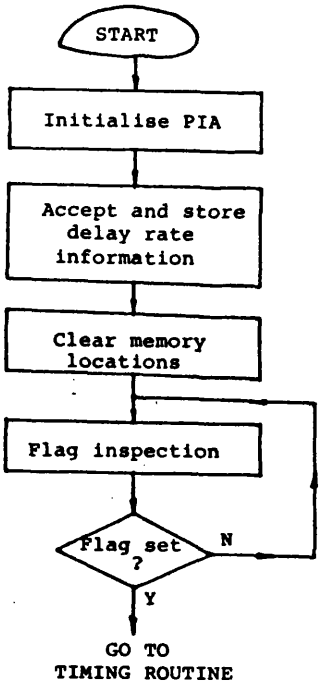


Figure A.4 The Initialise Routine

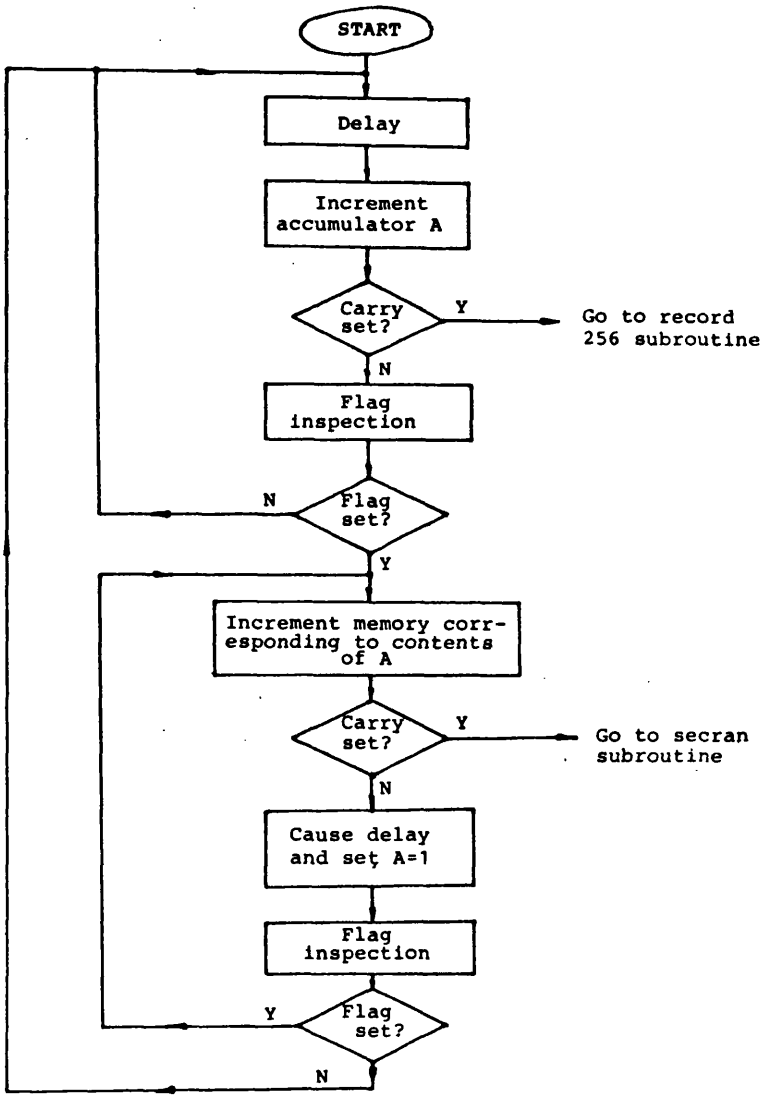


Figure A.5 The timing routine

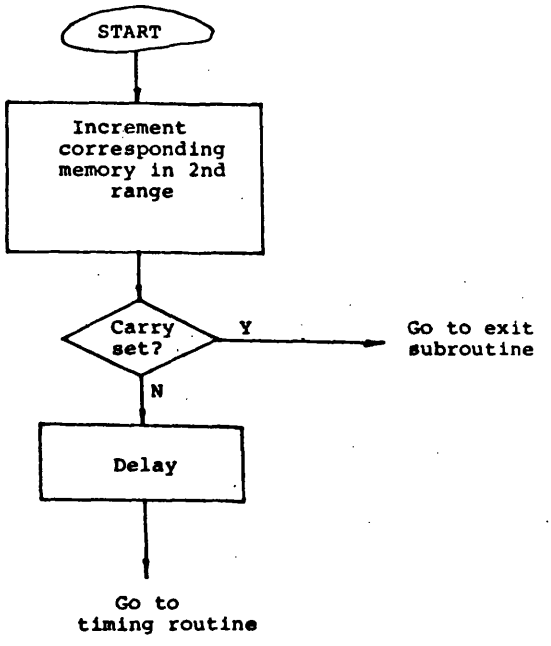


Figure A.6 The Secran Subroutine

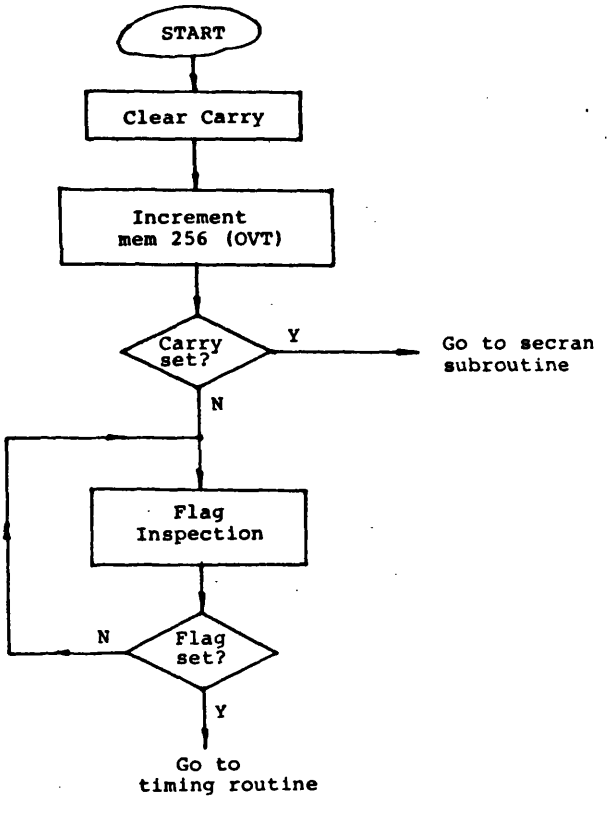


Figure A.7 The Record 256 Subroutine

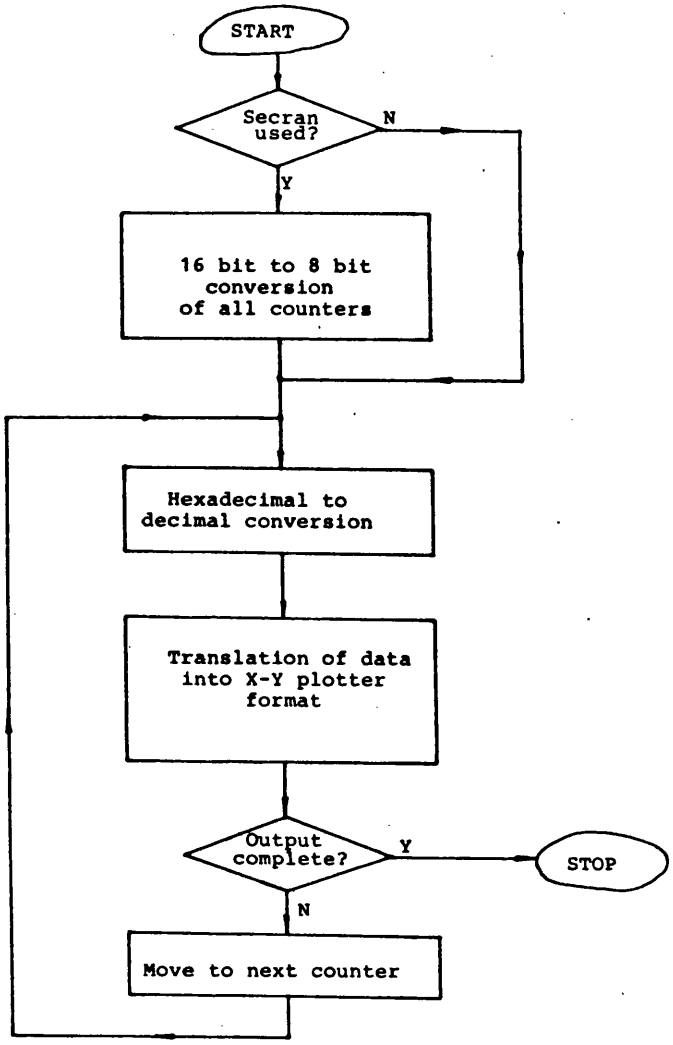


Figure A.8 The Exit Subroutine

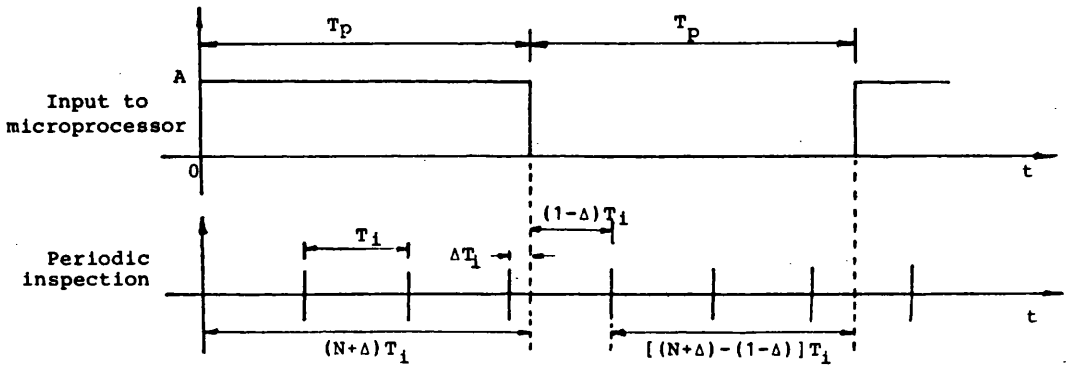


Figure A.9 The "Period" Timing Process

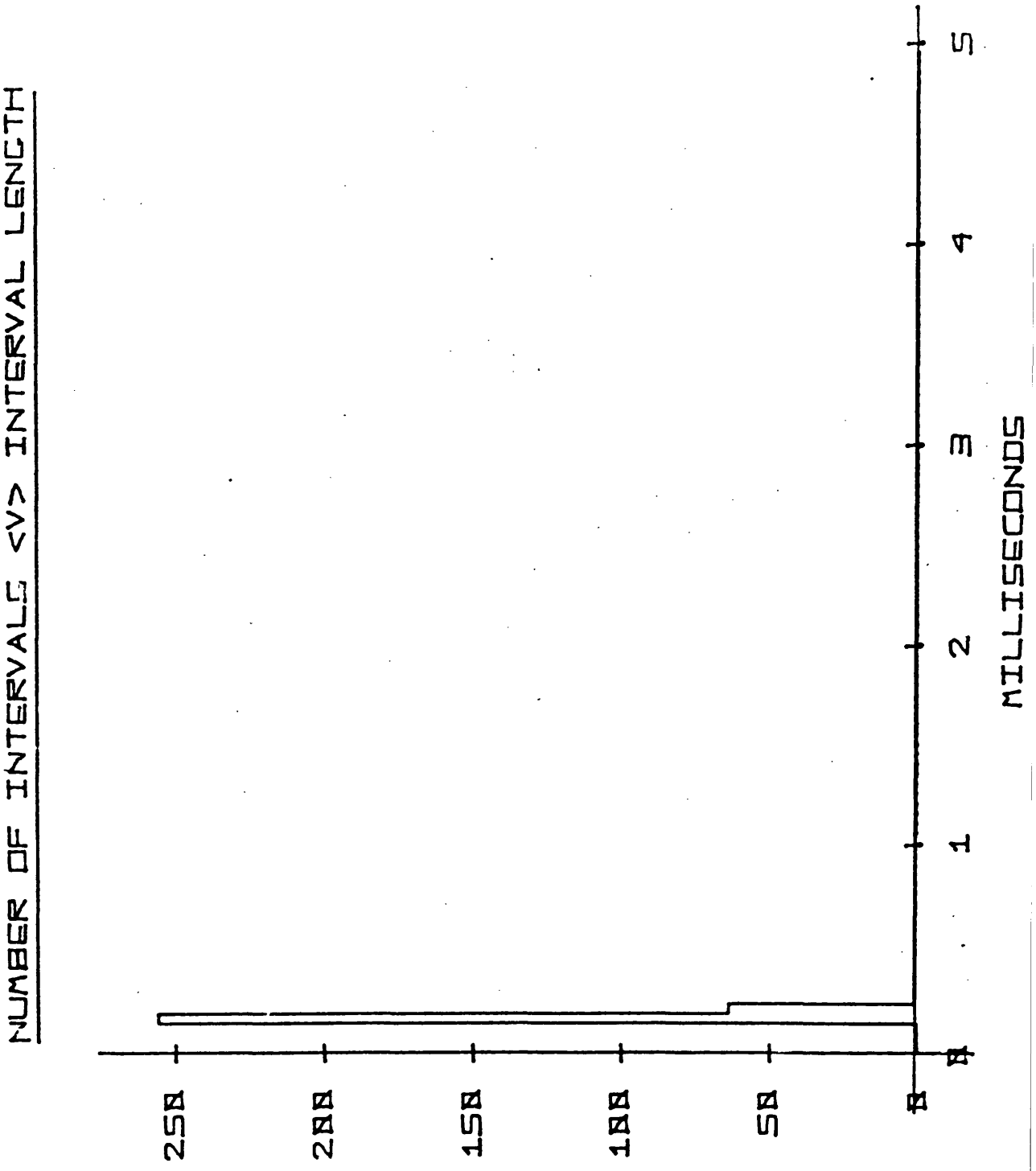


Figure A.10 The real-zero probability distribution for a sine wave with a period of 210μsecs